질의응답을 위한 복수문서 요약에 관한 실험적 연구*

An Experimental Study on Multi-Document Summarization for Question Answering

최 상 희(Sang-Hee Choi)**
정 영 미(Young-Mee Chung)***

초 록

이 연구에서는 이용자가 여러 곳에 분산되어 있는 문서들을 일일이 보지 않고 하나의 요약문에서 쉽게 질의에 맞는 답을 찾을 수 있는 가장 효율적인 방안을 제시하고자 하였다. 이를 위해, 클러스터링 기법, 단락확장 기법, 두 기법의 특성을 반영한 혼합 기법 등 세 가지 복수문서 요약 기법의 성능을 평가하는 실험을 수행하였다. 요약기법 평가 기준으로는 요약 정확률과 요약문내 정보 중복도를 적용하였다. 실험결과 이용자 질의에 따라 여러 문서를 요약하는 최적 기법으로 문장검색을 기반으로 한 순차적 단락확장 기법을 제안하였다. 순차적 단락확장은 특히, 요약의 대상이 되는 문서가 대용량인 환경에서 정확한 정보를 찾아 요약문을 생성하는 성능이 가장 우수한 것으로 나타났다.

ABSTRACT

This experimental study proposes a multi-document summarization method that produces optimal summaries in which users can find answers to their queries. In order to identify the most effective method for this purpose, the performance of the three summarization methods were compared. The investigated methods are sentence clustering, passage extraction through spreading activation, and clustering-passage extraction hybrid methods. The effectiveness of each summarizing method was evaluated by two criteria used to measure the accuracy and the redundancy of a summary. The passage extraction method using the sequential bnb search algorithm proved to be most effective in summarizing multiple documents with regard to summarization precision. This study proposes the passage extraction method as the optimal multi-document summarization method.

키워드: 복수문서 요약, 요약 기법, 문장 클러스터링, 단락확장, 질의응답
multi-document summarization, summarization method, sentence clustering, passage extraction, question answering

---

* 본 연구는 연세대학교 대학원 박사학위논문의 일부를 요약한 것임.
** 연세대학교 문헌정보학과 시간강사(shchoi@lis.yonsei.ac.kr)
*** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)
■ 논문접수일자 : 2004년 8월 27일
■ 게재확정일자 : 2004년 9월 13일

1. 서 론

자동요약은 문서검색, 웹 검색, 정보추출, 데이터마이닝, 질의응답 등의 분야에 적용되어 새로운 영역으로 발전하고 있다. 대용량 정보검색 환경에서 이용자가 적합한 정보를 찾기 위해 수백만 건에 달하는 문서를 읽어 볼 수 없기 때문이다. 따라서 효율적으로 원문 내용을 전달하는 방법이 필연적으로 요구되는 것이며 자동요약에 대한 요구도 다양해지고 있다.

이용자가 여러 문서에서 정보를 찾는 것은 정보요구를 해결할 수 있는 답을 찾기 위한 것이다. 답을 찾기 위해 일반적으로 이용자들은 질의에 따라 검색된 문서의 원문을 직접 다 보거나 또는 원문을 대표하는 요약문을 읽고 찾고 싶은 문서를 골라낸다. 이런 특성을 고려하여 여러 문서에서 이용자에게 찾고자 하는 부분만을 요약해서 제공하게 되면, 원문을 다 볼 필요가 없고 일부 중요한 부분만 필요한 이용자는 원하는 정보를 찾는데 시간과 노력을 절약하게 된다. 특히, 여러 문서에서 중복된 정보와 상이한 정보를 비교, 추출하여 제공하는 복수문서 요약 과정은 실제 정보검색 환경에서 이용자가 여러 문서에 흩어져 있는 질의에 대한 답이 되는 정보를 찾는 과정과 일치한다. 이용자는 특정 질의에 대하여 해답을 찾고자 할 때 여러 정보원에서 그 질의에 해당하는 부분만 수집하여 분석한 후 결합하여 사용하기 때문이다.

이 연구에서는 이용자가 여러 곳에 분산되어 있는 원문서를 일일이 보지 않고 요약문에서 쉽게 질의에 맞는 답을 찾을 수 있는 효율적인 방안을 제시하고자 한다. 이와 같은 목적으로, 클러스터링 기법, 단락확장 기법, 두 기법의 특성을 반영한 혼합 기법 등 세 가지 요약 기법의 성능을 평가하는 실험을 수행하여 복수문서 요약에 가장 효과적인 알고리즘을 제시하고자 하였다.

2. 이론적 배경

2.1 복수문서 요약과 요약 기법

최근 들어 다양한 연구가 이루어지고 있는 복수문서 요약은 일반적으로 이미 분류된 문서집단이나 검색 결과에서 포함된 문서들의 주제를 요약하는데 활용된다(Radeve, Jing, and Stys 2003). 이런 사례로 콜롬비아 대학의 뉴스기사 요약 시스템인 NewsBlaster는 여러 뉴스사이트를 돌아다니면서 뉴스를 검색하여, 검색된 많은 뉴스를 클러스터링한 후 요약하여 이용자에게 제공한다(Barzilay 2003). 또한, 통신사 뉴스처럼 같은 주제에 대해 연속적으로 새로운 정보가 발생할 때 정보발생 시점부터 가장 최근에 발생한 정보까지 추적하고 취합하여 하나의 요약으로 생성하는 연구도 있다. 최초 문서가 발생했을 때 기본 요약을 만든 후 추후 새로운 정보를 담은 문서가 입력될 때마다 비교하여 변경된 부분을 보완하고 새로운 정보를 추가하는 접근 방식이다(Radeve, Jing and Budzikowska 2000). 한 사건을 기준으로 시간차로 변경된 내용을 정리하는 복수문서 요약기능은 효과적인 방안이 될 수 있다(McKeown, and Radev 1999).

문서 자동요약을 생성하는 기법으로 적용되는 정보검색의 대표적 기법 중 하나는 문장 클러스터링이다(Radeve, Jing, and Budziko-

wska 2000, 정영미, 최상희 2001, Barizaly, Elhadad, and McKeown 2002, 장동현 2002). 문장 클러스터링은 문서를 구성하는 문장간의 통계적 유사성을 용어를 기반으로 산정하여 문장을 군집화함으로써 벡터 공간 상에서 가까운 거리에 있는 문장들은 같은 클러스터에, 그렇지 않은 문장들은 다른 클러스터에 포함시키는 일련의 분류작업을 거쳐 문서의 주제 구조를 탐색하는 것이다. 문장을 분류하여 각 클러스터에 같은 주제를 가진 문장들이 모이게 되면 클러스터로부터 요약문에 포함시킬 대표문장을 선별한다.

문장과 같이 정보검색 연구에서 많이 언급되며 요약에서도 주목을 하는 정보의 단위는 단락이다. 질의의 주제에 해당하는 단락을 개념확장 활성화 기법을 적용하여 동적으로 추출해서 요약문으로 생성하는 연구에서는 추출된 단락이 요약문의 기능을 하는 것으로 나타났다(김정하 2001). 이 연구에서 개념확장 활성화 기법으로 사용한 bnb (branch-and-bound search) 기법은 확장 중심노드에서 가장 유사한 방향으로 노드를 추적해가는 방식이므로, 특정 문장을 중심으로 가장 유사도가 높은 문장으로 확장해나갈 수 있는 방안을 제공할 수 있다.

2. 2 질의응답과 문서요약

일반적으로 질의응답시스템은 사실검색형 질의에 대한 답을 단답식으로 제공하는 시스템을 말한다. 최근 질의응답시스템에 대한 연구는 정보검색 분야와 정보추출 (information extraction) 분야에서 주목을 받게 되면서 좀 더 다양한 형태로 발전되고 있다. 예를 들면 질의에 대한 답이 특정 값(value), 명칭(name), 연도, 구 같은 짧은 형태에서 문장, 단락, 요약 같은 좀더 큰 정보단위로 확장되어 가고 있다. TREC-8에서는 질의에 답하는 정보단위로 단락을 추출하여 분류한 후 질의와 비교하여 제공하는 방법이 소개되기도 했고, TREC-9에서 시도된 질의응답 실험에서도 질의에 대한 답으로 50 또는 250 바이트에 해당하는 텍스트 조각(fragment)을 찾아내는 시도를 하였다(Clarke et al 2001).

질의응답 분야에서 새로운 대안을 제시할 수 있는 방법으로 복수문서 요약을 들 수 있다. 여러 문서에 흩어져 있는 정보를 수집, 요약한다는 복수문서 요약 개념을 질의응답 분야에 적용하였을 경우 산출된 요약문을 질의에 대한 답을 제공하는 방안으로 활용할 수 있기 때문이다. 또한, 이용자 중심의 요약 접근 방식을 적용하면 이용자가 실제 가지고 있는 정보요구인 요약의 주제란 질의응답시스템에서 질의의 개념에 상응한다고 할 수 있다. 이용자가 알고 싶어 하는 정보, 즉 질의를 중심으로 여러 문서에서 연관된 정보를 요약하여 충분히 제공함으로써 이용자가 원하는 정보를 요약 내에서 완전히 얻어내어 원문서를 더 이상 필요로 하지 않게 된다면 이용자가 가진 질의에 대하여 복수문서 요약이 답을 제공하게 되는 것이다.

3. 복수문서 요약 실험

3. 1 실험개요

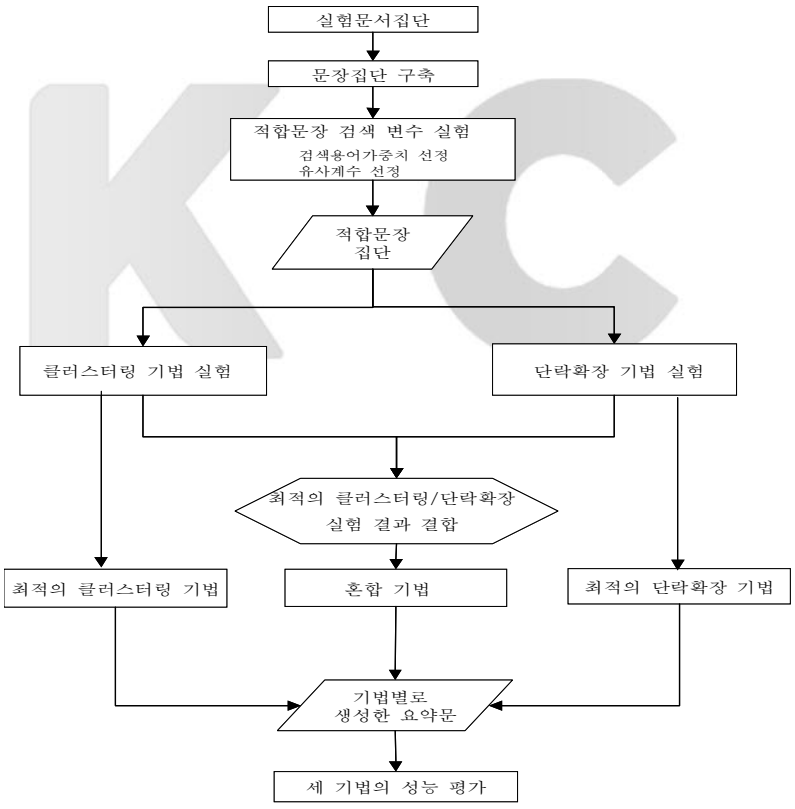
이 연구에서는 요약 기법을 적용하는 환경

을 복수문서와 질의중심 요약으로 규정하고, 요약 기법을 최적화하기 위해서 실험을 통해 이들의 성능과 특성에 영향을 주는 변수를 선정하였다. 실험 과정은 최적의 요약 알고리즘을 선정하기 위해 적합문장 추출을 위한 용어 가중치와 유사계수 성능 평가, 클러스터링 기법 평가, 단락확장 기법의 평가 분석을 위한 실험의 단계로 구성된다. 각 단계에서 수행된 실험 내용은 <그림 1>과 같다.

첫째, 요약의 선행단계로 복수문서에 흩어져 있는 질의와 관련된 정보를 검색, 추출하기 위하여, 질의-문장간 유사도 산출에 사용될 용어 가중치와 유사계수들의 성능 비교를 하는 실험을 수행하였다.

둘째, 클러스터링 기법으로 요약문을 생성하는 최적의 환경을 도출해내기 위해 클러스터링 기법 중 센트로이드 기법과 워드 기법을 택하여 문장을 클러스터링한 후 성능을 평가하였다. 클러스터링 기법의 요약 성능을 평가하는 방식으로 각 기법별로 생성된 클러스터의 크기를 비교하여 클러스터 크기의 편차를 살펴보았고 클러스터 크기의 편차가 요약문 생성에 미치는 영향을 분석하였다.

셋째, 질의의 답에 해당하는 정보를 추출하는 방식으로 문장벡터를 기반으로 한 검색 결과 중 질의와 유사도가 높은 문장 중심으로 단



〈그림 1〉 질의중심 복수문서 요약 기법 실험 과정

락을 확장하여 요약을 생성하는 기법의 성능을 비교하였다. 주제문장 중심의 단락확장 기법으로 순차적 bnb 활성화 기법을 적용한 순차적 단락확장 기법과 병렬적 bnb 활성화 기법을 적용한 병렬적 단락확장 기법을 평가하였다.

넷째, 클러스터링 실험과 단락확장 기법의 실험 결과 중 가장 좋은 성능을 나타낸 기법을 결합하여 혼합 기법을 실험하였다. 혼합 기법은 단락확장 기법에서 단락확장 기준이 되는 중심문장을 클러스터링 기법에서 추출한 대표문장 중에서 선정하는 방식이다. 혼합 기법으로 생성한 요약문은 각각의 단독 기법으로 생성된 요약문과 비교하여 질의에 적합한 정보를 제공하는 정확성과 중복성을 측정하였다.

3. 2 실험문서집단과 요약 알고리즘 성능 평가 방법

이 연구에서는 문서집단의 주제적 특성이나 구조적 특성에 독립적인 상황에서 요약문을 생성하는 기법을 연구하고자 하였다. 따라서 연구의 특성에 적합한 일반적인 주제를 다루고 있고 문서의 성격도 다양한 실험문서집단을 새로이 구축하게 되었다. 구축된 실험문서집단은 인터넷에 올려진 주간조선 기사 2,366건과 주간한국 기사 634건으로 총 3,000건이다. 이 중 주간한국 기사로 구성된 문서집단 600건은 요약알고리즘의 변수 선정을 위한 적합문장 추출 실험에서 실험집단으로 사용하였다. 실험문서집단의 문서 크기와 자수는 <표 1>과 같고 문장 당 평균 글자 수는 클러스터링 기법에서 생성하는 클러스터의 수를 고정하는데 반영되었다.

〈표 1 〉 실험문서집단에 대한 통계치

	문서당 문장 수	문장당 글자 수
평 균	134.7	83.4
최대값	211	246
최소값	45	8

이 연구에서 답을 제공하고자 하는 질의는 주제성검색형 질의와 사실검색형 질의의 중간 형태라고 할 수 있다. 사실검색은 보편적으로 단답형 형태의 답을 찾는 경우가 대부분이지만, 주제검색의 경우 검색의 시작과 끝이 불분명할 정도로 계속 심도 깊게 주변정보까지 검색하는 경우에서부터 한 장의 문서정도로 원하는 주제정보를 찾을 수 있는 경우까지 주제적 깊이와 필요로 하는 정보의 양이 다양하다. 여러 문서에서 문장단위로 정보를 수집하여 요약해서 제공하고자 하는 이 실험과 같은 환경에서 적절한 검색은 단답형의 사실을 원하는 검색보다는 적어도 몇 문장 이상의 정보를 원하는 주제검색의 성격이 있어야 한다.이 연구에서는 이와 같은 질의의 성격을 고려하여 정치, 사회, 경제, 문화 분야에서 고유명사와 일반명사를 혼합하여 요약을 생성할 질의 30개를 생성하였다. 요약문 생성 실험에 사용된 질의는 질의 당 평균 단어 수는 5.4개이다.

요약문의 품질을 평가하는 내재적 평가(intrinsic evaluation)는 요약문을 분석하여 요약 내 주요 개념이 포함되어 있는지, 요약이 읽기 쉬운지, 또는 사람이 수작업으로 만든 요약문과 비교를 하여 얼마나 유사한지 정도 등을 측정하여 자동으로 생성된 요약문의 품질을 평가하는 방식이다. 내재적 평가에 대응되는 개념인 외재적 평가(extrinsic evaluation)는 요약문이 내포하고 있는 정보의 양이나 질로

요약문의 품질을 평가하는 것이 아니라 요약문이 다른 일을 할 때 얼마나 효과적으로 이용될 수 있는 지를 조사하여 측정하는 평가방식이다. 이 연구에서는 두 가지 평가척도를 다 활용하였다.

내재적 평가척도로서 선정한 평가척도는 생성된 요약문내 적합한 정보가 포함되어 있는 비율을 측정한 요약 정확률과 중복되는 문장이 포함되어 있는 비율을 측정한 요약문장 중복률이다.

$$\text{요약 정확률} = \frac{\text{질의에 적합한 문장 수}}{\text{요약문 내 총 문장수}}$$

$$\text{요약문장 중복률} = \frac{\text{정보가 중복되는 문장 수}}{\text{요약문 내 총 문장수}}$$

외재적 평가를 하기 위해서 이용자가 요약문의 실제 활용성을 평가하도록 하였다. 요약문을 생성할 때 사용한 질의로부터 만든 문제를 이용자 열 명에게 주고 세 가지 요약 기법에 의해 생성된 요약문에서 해당하는 답이 될 만한 정보를 각각 찾도록 하여 질의응답 성능이 가장 좋은 요약 기법을 알아내고자 했다. 또한 이용자가 답이 되는 정보에서 중복되는 정보를 표시하도록 하여 요약문내의 응답정보 중복률을 측정하였다. 이용자에게 효과적인 요약문은 답이 되는 구는 많지만 그 안에 중복되는 구는 없는 요약문이다.

$$\text{응답정보 중복률} = \frac{\text{중복되는 답을 포함하는 구의 수}}{\text{요약문 내 답을 표현한 구의 수}}$$

일부 요약문의 경우 응답이 되는 구가 없다고 판명되는 사례가 있을 수 있으므로 응답이 되는 정보 수가 없는 경우를 응답실패 사례로 평가하였다. 응답실패율은 각 요약기법으로 생산된 요약문의 질의응답 성능을 측정한 것이다.

$$\text{응답실패율} = \frac{\text{응답되는 정보가 없다고 판정된 요약문의 수}}{\text{전체 요약문의 수}}$$

마지막으로 요약문에서 종합적 성능을 평가하기 위하여 이용자가 질의별로 답을 찾는데 가장 쉬었던 요약문 순으로 세 기법으로 생성된 요약문에 순위를 매기게 하였다.

3.3 적합문장 추출을 위한 용어가중치 및 유사계수 실험

이 실험에서 사용된 용어가중치는 문장 내 단어빈도를 이용한 세 가지 단어빈도 가중치와 단어빈도에 역문장빈도 가중치를 반영한 단어빈도*역문장빈도 가중치 세 가지 등 총 여섯 가지이다.

단어빈도 용어가중치로 문장 내 출현한 빈도를 가중치로 하는 단순단어빈도와 문장 내 출현여부만을 0, 1값으로 표현하는 이진(binary) 빈도, 문장 내 단어빈도의 영향력을 보완한 로그 단어빈도를 사용하였다. 로그 단어빈도는 단어빈도가 1인 용어의 지나치게 낮은 영향력을 보완하고 단어빈도가 높은 용어의 불필요한 영향력을 낮추기 위해 정보검색 실험 TREC-1에서 SMART팀이 제안한 공식이다.

$$\text{로그 tsf} = 1 + \log(\text{tsf})$$

단어빈도에 적용되는 역문장빈도(isf)는 스파크존스(Sparck Jones 1972)의 역문헌빈도를 응용한 것이다.

$$\text{isf} = 1 + log2\frac{NS}{sf}$$

NS : 문장 집단 내 문장의 총수
sf : 문장빈도

문장간 유사도 비교를 할 때 적용되는 유사도 공식 중 코사인계수와 내적유사도 계수를 사용하였다. 문장 x와 y가 벡터 길이 t일 때 각 문장의 용어i로 벡터간 유사도를 나타낸다면 다음과 같다.

$$\text{내적유사도(x, y)} = \sum_{i=1}^{t} x_i \cdot y_i$$

$$\text{코사인유사도(x, y)} = \frac{\sum_{i=1}^{t} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{t}(x_i)^2 \cdot \sum_{i=1}^{t}(y_i)^2}}$$

이 실험에서는 3개의 단어빈도 용어가중치와 각 단어빈도에 역문장빈도를 결합한 단어빈도 용어가중치 3개, 총 6개의 단어빈도 용어가중치에 내적유사도와 코사인 유사도 등 2개의 유사계수를 조합하여 12가지 조건에서 질의에 적합한 문장을 추출하는 성능과 검색결과 리스트에서 순위를 매기는 성능을 비교하였다. 그 결과 이진빈도와 역문장 가중치, 내적계수를 조합하여 적용한 것이 가장 성능이 좋은 것으로 나타났고 그 결과를 요약실험에 적용하였다.

3. 4 문장 클러스터링 기법 실험

적합문장으로 추출된 문장을 군집화하면 유사한 주제의 문장이 동일 클러스터에 모이게 되고, 각 문장 클러스터에서 질의와 가장 유사한 문장을 선정하면 질의와 적합하면서도 다양한 주제를 표현하는 문장들로 요약문을 생성할 수 있다. 이 연구에서는 가장 적절하게 처리할 수 있는 클러스터링 기법을 발견하기 위해 센트로이드 기법(centroid), 워드(Ward) 두 가지 방식의 클러스터링 기법을 비교하였는데, 그 이유는 문장 클러스터링을 위한 기법을 비교한 기존 요약 연구에서 센트로이드 기법과 워드 기법이 안정적인 것으로 나타났기 때문이다(정영미, 최상희 2001).

두 기법 비교를 위해서 질의 10개를 선정하여 적합문장집단을 구축한 후 각 문장집단에서 생성 클러스터 수를 6개로 고정한 후 클러스터를 생성하여 분석하였다. 클러스터의 수를 6개로 고정한 것은 이 실험에서 최종적으로 산출하고자 하는 요약문이 500자 요약문이기 때문이다. 즉, 실험문서집단의 문장당 평균 글자 수가 83.4자이므로 500자 요약문은 평균 6개의 문장으로 구성된다.

클러스터링 기법에서는 각 클러스터당 대표문장을 1개씩 선정하여 요약문을 생성하게 되므로 요약문에 포함되는 문장 수로 생성되는 클러스터의 수를 고정하게 된 것이다. 이 실험에서 평가 척도로는 클러스터 표준편차, 2문장 이하 소규모 클러스터 생성, 각 클러스터링 기법으로 생성된 요약문의 요약 정확률, 요약문장 중복률을 이용하였다.

먼저 클러스터를 생성하는 측면에서는 <표 2>의 표준편차 값에서 나타났듯이 워드 기법이 클러스터 크기를 고르게 생성하는 것으로 나타났고 문장 1개나 2개로 한 클러스터가 생성되는 2문장 이하 소규모 클러스터를 생성하는 경우도 없는 것으로 나타났다. 대표문장을 추출하기 위해 클러스터를 적용하는 문장 클러스터링에서 클러스터의 크기가 중요한 이유는 클러

스터의 크기가 크게 차이가 나게 형성될 때, 한 클러스터 내 각 문장들을 대표문장으로 선별하기 위한 환경이 공평하게 조성되지 않기 때문이다. 최악의 경우 주요문장 대부분이 한 클러스터에 몰리고 의미 없는 문장이 단일멤버 클러스터를 형성하게 되면 주요문장은 대표문장으로 선정될 기회가 최소화되고 어느 클러스터에도 속하지 못하여 따로 떨어지게 되는 무의미한 문장이 대표문장으로 선정될 기회가 최대화된다. 그러므로 클러스터 크기 편차가 크거나 단일클러스터를 많이 생성시키는 기법은 대표문장을 추출하려는 목적에 부합하지 않는다.

〈표 2〉 문장 클러스터링 기법의 성향 평가

	클러스터 크기 표준편차	2문장 이하 소규모 클러스터 생성 수
센트로이드	137.7	36
워 드	49.39	0

클러스터링 기법의 요약 성능은 센트로이드 기법과 워드 기법으로 생성된 클러스터의 대표문장으로 구성된 요약문의 정확도와 중복도를 기반으로 평가하였다. 요약문에 포함되는 각 클러스터의 대표문장은 클러스터내의 문장과 질의간의 유사도를 비교하여 질의와 가장 유사도가 높은 문장이다. <표 3>에서 보면 워드 기법이 요약 정확률에서 센트로이드 기법보다는 월등히 좋은 것으로 나타났다. 이는 <표 2>에서 나타났듯이 작은 클러스터가 많이 생성되면, 주제성이 모호한 문장들이 작은 클러스터에서 별다른 경쟁 없이 대표문장으로 선정되어 요약문으로 선정되면서 요약 정확률이 떨어지는 현상을 나타냈기 때문이다.

요약문장 중복률은 센트로이드가 낮은 것으로 나타났으나 이 현상 역시, 작은 클러스터의 영향으로 발생한 결과라고 분석된다. 클러스터링 기법 실험에서는 클러스터 편차, 2문장 이하 소규모 클러스터 생성, 요약 정확률 면에서 워드 기법이 센트로이드 기법보다 좋은 성능을 보였다.

〈표 3〉 문장 클러스터링 기법의 요약 성능 평가

요약문 생성 기법	요약 정확률	요약문장 중복률
센트로이드	0.183	0.016
워 드	0.416	0.094

3. 5 단락확장 기법 실험

단락확장의 방법으로 순차적 bnb 활성화 기법과 병렬적 bnb 활성화 기법을 적용한 순차적 단락확장 기법과 병렬적 단락확장 기법을 비교하였다.

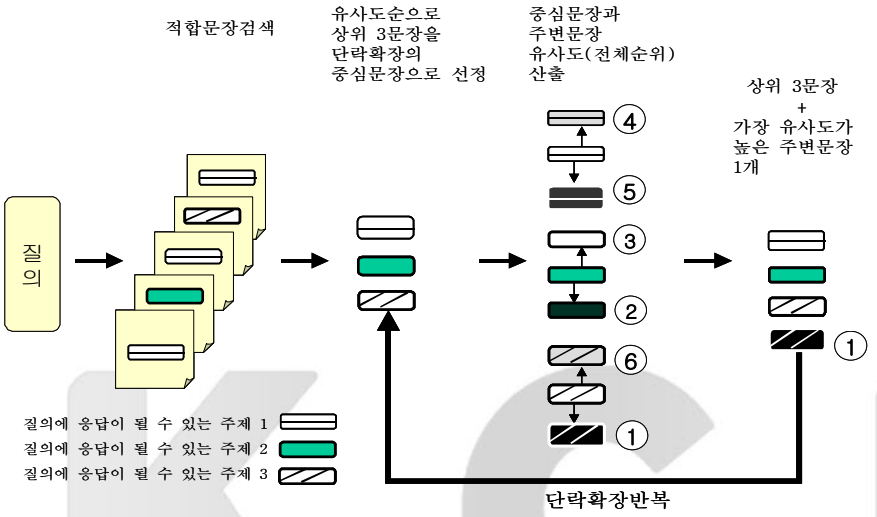
순차적 단락확장 기법은 주변단락으로 확장할 때 한 노드에서 다음 노드까지의 경로를 기억시킨 후 그중 최단 경로 1개만을 선택하여 확장하는 방식이다. 이때 포함된 최단경로는 다음 확장에 포함되며 2차 확장을 위해 다른 경로와 비교할 때 다시 최단경로를 선택하는 순환과정을 통해서 단락확장이 이루어지는 방식이다. 순차적 단락확장 기법을 적용하여 단락이 확장되는 과정은 <그림 2>와 같다.

<그림 2>에서처럼 단락확장이 실제 문서에 적용되는 과정을 살펴보자면 순차적 단락확장의 첫 단계는 검색결과의 1, 2, 3위의 문장을 중심으로 앞뒤 단락들과 유사도를 산출하여 유사도가 큰 순으로 순위를 매기게 된다. 순차적 단락확장 기법에서는 중심 문장과 비교한 앞뒤 단락의 유사도를 모두 한 곳에 모은 다음 유사

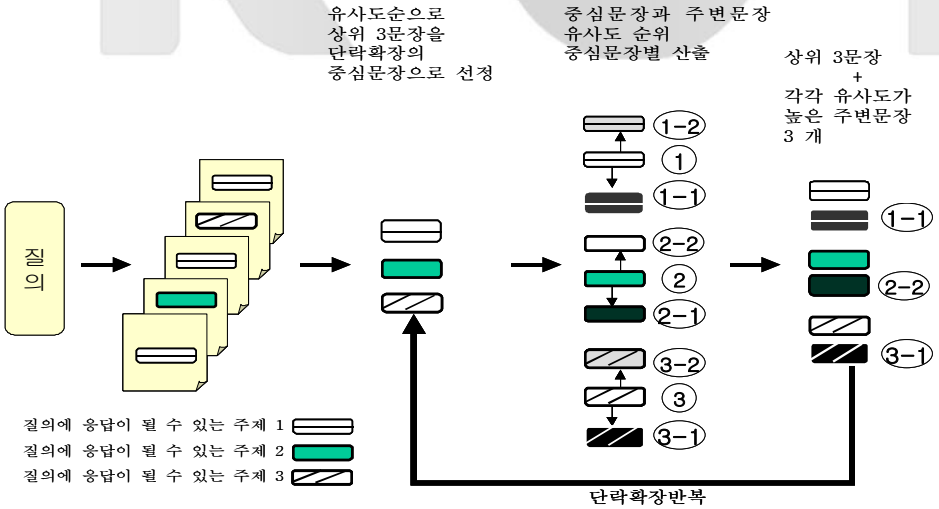
도 순으로 전체 순위를 매기게 되므로 비교된 전체 단락 중 가장 유사도가 높은 단락이 추출된다. 따라서 요약문에 포함시킬 단락을 최종적으로 추출할 때 중심 문장과의 각각 개별적인 관계가 직접적으로 반영되는 것은 아니다.

병렬적 단락확장은 순차적 단락확장이 시작노드에서 가장 최단 경로만을 포함시켜 확장하는 방식을 택하는 것에 반해 시작노드에서 확장된 노드 각각에 대하여 2차 확장이 이루어지는 것이다. 병렬적 단락확장 과정은 <그림 3>과 같다.

병렬적 단락확장이 문서에 적용되는 과정을 살펴보자면, 질의와 상의 세 문장을 단락확장에서 중심 문장으로 하여 주변 단락과 유사도



〈그림 2〉 순차적 단락확장 알고리즘



〈그림 3〉 병렬적 단락확장 알고리즘

를 비교하는 것은 순차적 단락확장의 첫 단계와 같다.

그 다음 단계로 단락을 병렬적으로 확장하는 과정은 1, 2, 3 위의 문장을 중심으로 앞뒤 단락들과 유사도를 산출하여 각 중심 문장별로 유사도가 큰 순으로 순위를 매겨 중심문장 당 1개씩 유사도가 높은 단락을 추출하는 것이다. 따라서 병렬적 단락확장 기법으로 주변 단락을 추출할 때는 순차적 단락확장 기법과는 달리 중심문장과 주변 단락의 개별적인 관계가 직접적으로 반영된다.

이 실험에서 적용한 순차적, 병렬적 단락확장 기법은 기존 연구(김정하 2001)에서 사용된 기법을 기반으로 하였다. 그러나 여기서 소개된 기법은 단일문서 환경에서 적용된 것이므로 동일한 단락확장 중지 규칙을 적용한 결과, 복수문서 환경에서는 단락확장 중지가 예상보다 빈번하게 발생하는 문제점이 발생하게 되었다. 따라서 요약문의 기본 정보량을 충족시키지 못하고 단락확장이 중지되었을 때 해결방안으로 중심문장을 질의와 유사한 순으로 추가해 가도록 단락확장규칙을 보완하여 적용하였다.

<표 4>에서 나타났듯이 성능 평가 결과 순차적 단락확장 기법이 병렬적 단락확장 기법보다 질의에 적합한 문장을 효과적으로 추출하였다. 순차적 단락확장 기법으로 추출한 요약문에 적합한 문장이 포함되어 있는 비율은 0.638로 병렬적 단락확장 기법의 0.561 보다 13% 높았다

요약문내 문장들이 유사한 정보를 중복하여 제공하는 비율도 순차적 단락확장이 더 낮은 것으로 나타났다. <표 4>를 보면 순차적 단락확장 기법이 요약 정확률은 상대적으로 높게 유지하면서 낮은 중복률을 나타내고 있다. 병렬적 단락확장 기법은 적합정보를 추출할 때 같은 주제를 집중적으로 추출해낸다는 성향을 지니고 있다는 것이 판명되었다. 반면 순차적 단락확장 기법은 유사한 주제로 확장해나가지만 상대적으로 주변 주제로 확장해나가는 성향이 있다. 단락확장 실험에서는 순차적 단락확장 기법이 요약 정확률도 유지하면서 질의에 적합한 정보를 상대적으로 적게 중복되도록 요약문에 포함시키는 것으로 평가되었다.

〈표 4〉 단락확장 기법의 요약 성능 평가

단락확장 기법	요약 정확률	요약문장 중복률
순차적 단락확장	0.638	0.219
병렬적 단락확장	0.561	0.330

3.6 문장 클러스터링과 단락확장 혼합기법 실험

단락확장 기법에서는 단락확장의 시발점이 되는 중심문장들을 적합문장 검색 결과에서 선정하는 것이 가장 중요한 단계이다. 질의에 대한 정보를 추출하기 위해서 질의와의 유사도가 가장 우선되는 기준이 되겠지만 이 경우 중심문장간의 유사도는 무시된다. 즉, 질의와의 유사도를 측정해 상위 3개의 문장을 중심문장으로 선정한다고 했을 때, 1위의 문장과 2위, 3위의 문장이 거의 같은 주제를 담고 있는 문장이라면 3개의 문장을 가지고 확장을 하더라도 결국 하나의 주제로 단락확장을 하는 결과가 된다. 최악의 경우 1개의 문장으로 확장한 결과와 제공하는 정보는 비슷해질 수 있다. 따라서 중심문장간의 유사도가 측정되지 않으면 제공되는 정보가 중복되는 비율이 높아지는 결과를 산출할 수 있다.

클러스터링 기법을 적용한다면 상위문장을

중심으로 하는 단락확장시 주제의 중복성을 최소화할 수 있다. 즉, 중심문장의 주제적 상이성을 확보하기 위해 단순히 적합문장 검색 결과 상위순위의 문장을 이용하는 것이 아니라 클러스터링을 해서 산출된 대표문장을 단락확장의 중심문장으로 적용하는 것이다. 클러스터의 대표문장은 클러스터링 생성시 주제의 유사성이 비교 측정된 것으로 다른 클러스터에 속하여 있다는 것은 어느 정도 주제가 다르다는 것을 보장한다.

혼합 기법은 클러스터링에서 추출된 대표문장 중 질의와 유사도 순으로 상위 3개 문장을 가지고 순차적으로 단락확장을 하는 요약문을 생성하는 방식이다. 혼합 기법은 이전의 각 기법별 실험 단계에서 최상의 성능을 보인 결과인 워드 기법과 순차적 단락확장 기법를 결합하여 생성하였다.

클러스터링 기법, 단락확장 기법, 클러스터링 - 단락확장 혼합 기법으로 산출된 요약문의 성능은 <표 5>와 같다. 클러스터링 기법, 단락확장 기법, 혼합 기법을 비교하면 요약 정확률은 순차적 단락확장 기법이 0.638로 정확한 문장을 요약문에 포함시키는 성능이 가장 좋은 것으로 밝혀졌고 클러스터링 기법은 0.416으로 가장 낮았다. 반대로 요약문장 중복률에 있어서는 클러스터링 기법이 0.094로 중복되는 정보가 가장 없는 것으로 나타났고 단락확장 기법이 0.219로 가장 높았다.

결합된 기법들의 특성을 적용하여 분석하면 혼합 기법의 요약 정확률은 0.620으로, 요약 정확률이 0.416으로 나오는 클러스터링 기법의 특성이 단락확장 기법으로 인해 보완된 것으로 나타났다. 또한 요약문장 중복률 측면에서도 혼합 기법의 0.167은 단락확장 기법의 0.219보다 개선된 결과를 나타내었다

〈표 5〉 혼합기법과 다른 기법의 요약 성능 비교

요약문 생성 기법	요약 정확률	요약문장 중복률
혼합 기법	0.620	0.167
클러스터링 기법	0.416	0.094*
단락확장 기법	0.638*	0.219

*가장 성능이 좋은 결과

4. 요약 기법의 질의응답 성능 평가

생성된 요약문의 실제 활용도를 평가하고자 30개의 질의에 따라, 각 기법실험에서 최적의 결과를 나타난 세 가지 기법별로 생성된 90개의 요약문을 무작위로 선정된 문헌정보학과 대학원생인 10명의 이용자에게 보여주고 각 요약문에서 질의에 답이 되는 정보를 체크하라고 하였다. 이는 실제 요약문이 질의에 응답할 수 있는 정도를 측정하는 것으로, 답이 되는 정보의 단위는 맥락이 이어지는 구로 선정하였다.

〈표 6〉 요약문의 질의응답 성능 평가

	단락확장	클러스터링	혼합
답이 되는 정보의 평균 수	6.3	4.5	6.1
응답정보 중복률	0.416	0.208	0.375
응답실패율	0.033	0.080	0.036

평가척도는 3.2에서 설명한 외재적 평가척도를 적용하였으며 그 결과는 <표 6>에서 나타났듯이 답이 되는 정보를 많이 포함하고 있다고 판정된 요약 기법은 단락확장 기법이다. 이는 요약알고리즘 평가에 사용된 요약 정확률

결과와도 상응하고 있고, 실제 이용자가 정확한 정보가 많이 포함되었다고 판단한 단락확장 기법은 응답실패율도 낮은 것으로 나타났다. 중복되는 정보를 평가하는 과정에서는 클러스터링 기법이 가장 중복도가 낮은 것으로 판명되었고 요약 정확률이 비슷하게 나타난 단락확장 기법과 혼합 기법 두 기법 중에는 혼합 기법이 단락확장 기법보다는 중복률이 낮은 것으로 나타났다.

이와 같은 요약 정확률과 응답실패율, 답이 되는 정보의 수로 나타나는 기법간의 차이가 실제 이용자에게 어느 정도로 느껴지고 있는지를 파악하기 위하여, 각 질의별로 답을 찾는데 가장 효과적이었던 요약문 순으로 순위를 매기도록 하였다. 순위를 선정할 때 동률 순위도 허용하여 같은 순위로 매겨진 경우도 발생하였고 그 결과는 <표 7>과 같다.

〈표 7〉 세 기법의 질의응답 성능 비교

	단락확장	클러스터링	혼합
평균 순위	1.496	2.460	1.630
1등한 횟수	180	37	142

평균 순위는 기법별로 판정받은 순위의 평균을 낸 것으로 수치가 1에 접근할수록 이용자가 1위로 평가한 경우가 많은 것이다. <표 7>에서 보여지듯이 이용자는 순차적 단락확장을 가장 좋은 요약문으로 선택한 경우가 제일 많았고 클러스터링 기법을 가장 성능이 낮은 것으로 판단하였다. 클러스터링 기법의 질의에 대한 응답 성능이 낮게 판명된 것은 요약 정확률 측면에서 너무 낮은 성능을 보였기 때문에 기인한 것으로 분석된다. 그러나 혼합 기법이 단락확장과 정확도면에서 유사한 성능을 보였고 다양성에서는 나은 성능을 보였음에도 불구하고 단락확장 기법만큼 좋은 평가를 받지 못한 것은 실제 이용자들이 요약문내 정보의 다양성보다 정확한 정보가 중복되더라도 많이 나타나는 것을 선호한다는 사실로 분석될 수 있다.

5. 결 론

각 단계별로 수행된 실험과 요약문 평가에서 도출된 결과는 다음과 같다.

첫째, 질의와 문장을 비교하여 적합문장을 추출하는 실험에서는 역문장단어빈도*이진단어빈도 용어 가중치와 내적계수를 조합한 결과가 가장 좋은 성능을 보였다. 단어빈도에 역문장빈도를 적용한 용어가중치가 단어빈도만을 적용한 용어가중치보다 성능이 좋게 나온 것은 문장처럼 저빈도어가 일반적인 환경에서 역문장빈도와 같은 저빈도어 가중치를 보정해주는 용어가중치가 효과적이기 때문이다.

둘째, 요약문을 생성하는 클러스터링 기법 실험에서는 워드 기법이 센트로이드 기법보다 성능이 좋은 것으로 나타났다. 클러스터링 기법의 요약 정확률은 0.416으로 단일문서 환경에서 문장 클러스터링을 적용하였던 이전 연구(정영미, 최상희 2001)에서 나타난 결과 0.47과 비교해도 떨어지는 수치로 복수문서 환경에서 문장 클러스터링이 요약문 생성에 적합한 기법이 아닐 수 있다는 것을 보여주었다.

셋째, 단락확장 기법 실험에서는 순차적 단락확장 기법과 병렬적 단락확장 기법을 비교한 결과 순차적 단락확장 기법이 적합한 정보로 확장하는 정확도나 확장된 정보 내에 유사정보

가 포함되는 정도인 중복도 측면에서 모두 성능이 좋은 것으로 나타났다. 단락확장 기법의 요약 정확률 0.638은 단일문서 내에서 순차적 단락확장 기법과 병렬적 단락확장 기법을 적용한 이전 연구(김정하 2001) 결과인 0.510(순차적 단락확장 기법), 0.505(병렬적 단락확장 기법)보다 높은 것으로 나타나 전반적으로 단락확장 기법은 복수문서 환경에서 성능이 향상된 것으로 나타났다.

넷째, 혼합 기법 실험에서는 각각 단독으로 클러스터링 기법과 단락확장 기법, 클러스터링-단락확장 혼합 기법으로 생성된 요약문을 비교하여 혼합 기법의 특성을 분석하였다. 혼합 기법은 요약 정확률 0.620으로 단락확장 기법보다는 다소 낮은 편으로 나타났다. 이는 클러스터링 기법에서 선정된 대표 문장을 단락확장의 중심문장을 사용할 경우 질의중심으로 검색된 상위 세 문장으로 단락확장의 중심문장을 선정하는 단락확장 기법보다 정확률이 떨어지기 때문이다. 그러나 요약문장 중복률을 보면 혼합 기법이 단락확장 기법의 중복률보다 낮은 것으로 평가되어 요약문의 정확도는 크게 저하시키지 않으면서 요약문장 중복률은 낮추는 기능을 하는 것으로 나타났다. 요약문장 중복률 측면에서 보면 클러스터링 기법이 가장 낮은 수치를 보이고는 있으나 요약 정확률 측면에서 현저하게 낮은 성능을 보이고 있어 복수문서 요약문 생성 기법으로 적합하지 않은 것으로 보인다.

다섯째, 이용자가 세 가지 기법을 생성된 요약문을 평가한 결과, 답이 되는 구의 평균 수에서는 단락확장 기법, 혼합 기법, 클러스터링 기법 순으로 많은 응답정보를 제공하는 것으로 평가되었다. 응답실패율도 단락확장 기법의 성능이 0.033으로 나타나 0.080으로 결과가 나타난 클러스터링 기법과는 크게 차이가 나며 혼합 기법보다 낮은 것으로 나타났다. 반면 응답정보 중복률에서는 클러스터링 기법, 혼합 기법, 단락확장 기법 순으로, 클러스터링 기법이 가장 중복되는 정보를 적게 요약문을 생성하는 것으로 나타났다. 세 기법의 평가결과를 종합하면 가장 정확한 정보를 제공하는 요약 기법은 단락확장 기법이며 가장 안정적인 성능을 보인 것은 혼합 기법인 것으로 나타났다.

여섯째, 세 가지 요약 기법에 따라 생성된 요약문의 순위 평가에서는 단락확장 기법이 혼합 기법보다 좋은 순위로 선정된 것으로 나타났다. 평균 순위에서 단락확장 기법이 1.496이고 혼합 기법은 1.630이므로 결과적으로 이용자가 단락확장 기법을 세 기법 중 1등으로 선정한 경우가 많은 것으로 나타났다. 이는 혼합 기법의 요약문장 중복률이 낮아졌음에도 불구하고 요약 정확률이 단락확장보다 높지 않았기 때문에 나타난 결과라고 분석된다. 이용자는 질의에 응답이 되는 정보를 요약문에서 찾는데 있어서 정보의 중복성보다 정확성에 더 영향을 많이 받는 것으로 나타났다.

이상의 실험결과를 통해서 이 연구에서는 복수문서 환경에서 이용자 질의에 따라 요약을 하는 가장 효율적인 방안으로 문장검색을 기반으로 한 순차적 단락확장 기법을 제안한다. 순차적 단락확장은 복수의 문서들과 같이 요약의 대상이 되는 정보가 대용량인 환경에서 정확한 정보를 찾아 요약문에 포함시키는 성능이 가장 뛰어난 것으로 나타났다. 또한 단락확장을 하는 기준인 문장을 선출해내는 단계에서 중복성

을 검증하는 기법을 결합한 혼합 기법은 중복성을 줄여주면서도 정확성은 크게 떨어뜨리지 않아 새로운 요약 기법으로의 가능성을 보여주었다.

참 고 문 헌

김정하. 2001. 『이용자 중심 요약문 생성에 관한 실험적 연구』. 석사학위 논문, 연세대학교, 문헌정보학과.

장동현. 2002. 『문장 클러스터링을 통한 텍스트 자동요약에 관한 연구』. 박사학위논문, 충남대학교, 컴퓨터 공학과.

정영미, 최상희. 2001. 문장 클러스터링에 기반한 자동요약 모형. 『정보관리학회지』, 18(3): 159-177.

정영미. 1993. 『정보검색론』. 개정판. 서울: 구미무역.

Barizilay, Regina, Elhadad, Noemie and Mckeown, Kathleen. 2002. "Inferring Strategies for Sentence Ordering in Multidocument News Summarization". *Journal of Artificial Intelligence Research,* 17: 35-55.

Barizilay, Regina. 2003. *Information Fusion for Multidocument Summarization: Parephrasing and Generation.* Phd. diss. Columbia University.

Callan, J. P. 1994. "Passage-level Evidence in Documentation Retrieval". *In Proceedings of the 17th Annual International ACM-SiGIR Conference on Research and Development in Information Retrieval,* 302-309.

Clark, C. L. A., Cormack, G. V., Kisman, D. I. E and Lynam. T. R. 2001. "Question Answering by Passage Selection." <http://citeseer.ist.psu.edu/454871.html>

Kaszkiel, M. and Zobel, J. 2001. "Effective Ranking with Arbitrary Passages". *Journal of the American Society for Information Science and Technology,* 52(4): 344-364.

McKeown, K. and Radev, D. 1999. "Generating summaries of multiple news articles" In *Advances in Automatic Text Summarization*. Cambridge: MIT Press

Radeve, D. R., Jing, H. and Budzikowska, M. 2000. "Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-based Evaluation and User Studies". In PANLP/NAACL 2000 Workshop, 21-29.

Radeve, D. R., Jing, Hongyan, Stys,

Malgorzata and Tam, Daniel. 2003. "Centroid-based Summarization of Multiple Documents." <http://cs-tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrl.cornell/TR94-1438>.

Sparck Jones, Karen. 1972. "A Statistical Interpretation of Term Specificity and Its Application in Retrieval". *Journal of Documentation,* 28(1): 11-20.