단어의 공기 관계 그래프를 이용한 문서의 핵심 문장 추출에 관한 연구

류 제[†] · 한 광 록[††] · 손 석 원[†††] · 임 기 욱[††††]

요 약

본 논문은 문서의 내용을 요약하기 위한 방법으로서 단어의 공기 관계 그래프를 이용한 핵심 문장 추출 방법을 제안한다. 문서에서 단어의 공기 관계 그래프를 이용하여 개념클러스터를 생성하고 문서내의 저자의 의도에 해당하는 주장을 찾는다. 그리고, 주장과 개념클러스터와의 관계로부터 키워드를 추출한다. 마지막으로 추출된 키워드와 주장을 이용하여 문서의 핵심 문장을 선택한다. 실험 및 평가는 수작업으로 추출한 핵심 문장과 비교를 통하여 이루어 졌으며, 기존의 방법과 비교하여 약 10%정도 향상된 성능을 보였다.

A Study on Automatic Extraction of Core Sentences from Document using Word Cooccurrence Graph

Ryu Je[†] · Kwang-Rok Han[††] · Seok-Won Sohn[†††] · Kee-Wook Rim[††††]

ABSTRACT

In this paper, we propose a method of core sentences extraction using word cooccurrence graph in order to summarize a document. For automatic extraction of core sentences, we construct a mean cluster from word cooccurrence graph, and find insistence which corresponds a purpose of author. And then we extract keywords by using relationship between mean cluster and insistence. Finally, core sentences are selected based on keywords and instances. The results are evaluated by comparing with manual extraction, and show that the extraction performance is improved about 10%.

1. 서 론

현대를 흔히 정보화 사회라고 한다. 우리는 쏟아지는 정보의 홍수 속에서 살고 있으며, 이미 우리 스스로가 처리할 수 있는 수준을 넘어선지 오래다. 특히, 인터넷과 같은 정보 유통 시스템의 발달로 인해 정보의 양은 하루가 다르게 지속적으로 증가하고 있으며, 이런 상황에서 제한된 시간에 누가 더 많은 정보를 얻느냐 하는 것은 중요한 문제로 대두되고 있다. 이러한 이유로 현재 많이 이용되어지는 인터넷 서비스 중에는 문서의 검색 및 분류 서비스가 상당수를 차지하고 있다. 그러나, 이러한 일반적인 검색 사이트들에서 검색 결과로서 제공되어지는 문서에 대한 설명이 부적절하거나 미약하며, 이로 인해 사용자가 원하는 문서를 찾기 위해서는 검색 결과로 제공되어지는 문서들은 사용자가 하나씩 읽어 보면서 확인하기에는 너무 많은 양이다. 따라서 정보 과적재(Information Overload)문제는 정보 검색 시스템에서 해결해야할 과제로 남아 있다[1]. 이에 이러한 웹 사이트에서 제공되어지는 문서의 내용을 간략하게 요약하여 검색 결과로서 함께 제

† 준 회 원 : 호서대학교 벤처전문대학원
†† 종신회원 : 호서대학교 벤처전문대학원 교수
††† 정 회 원 : 호서대학교 컴퓨터공학과 교수
†††† 종신회원 : 선문대학교 산업공학과 교수
논문접수 : 2000년 6월 19일, 심사완료 : 2000년 10월 4일

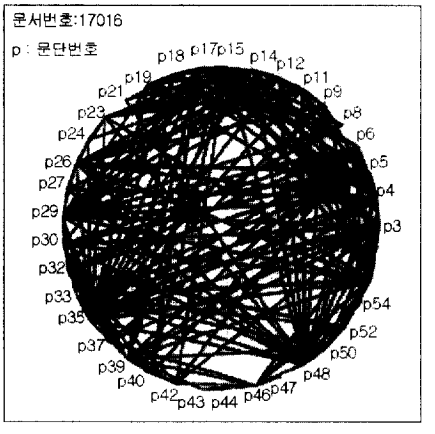
공되어지는 것은 필수불가결한 요소가 되었다[2]. 이렇듯, 문서 요약은 현대와 같은 정보화 사회에서 원하는 정보를 쉽게 찾거나 혹은 얻기 위해서 반드시 필요한 분야로서 자리를 잡고 있다.

문서 요약이란 문서의 기본적인 내용을 유지하면서 문서의 복잡도, 즉 문서의 길이를 줄이는 작업이다. 문서 요약에서는 요약문을 생성하는 방식에 따라 추출과 요약으로 나눌 수 있다[3]. 본 논문에서는 문서의 내용을 요약하기 위한 방식으로 핵심 문장 추출 방식을 사용하고 있으며, 문서에서 핵심 내용을 추출하기 위해서 우선 문서의 내용 중에서 키워드를 추출하는 방식을 사용한다. 본 논문에서는 키워드를 좀더 효과적으로 추출하기 위하여 단어간의 공기 관계를 이용한 방식을 키워드 추출 방법에 적용하였다. 그리고 본 논문에서는 시스템의 처리 속도 보다는 정확하게 관련된 문장을 추출하는데 중점을 두었다.

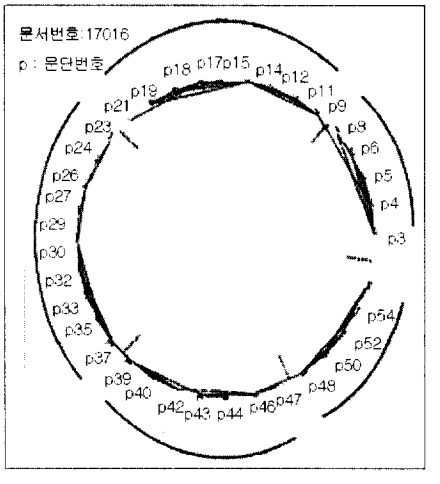
본 논문은 다음과 같이 구성된다. 우선 2장에서 기존의 문서 요약 방법에 대해서 살펴보고, 3장에서는 본 논문에서 제안하는 시스템 모델 및 핵심 문장 추출 과정을 살펴보며, 4장에서는 실험 및 평가를 그리고 5장에서 결론을 맺는다.

2. 관련연구

요약을 한다는 것은 우선 문장에 대한 이해, 중요한 주제의 선별, 요약문 생성의 세단계로 크게 볼 수 있다[4]. 특히 요약문의 생성 방법에 따라 크게 추출과 요약으로 그 방법을 나눌 수 있는데, 최근 들어서는 추출에 관련된 연구가 많이 이루어 지고 있다. 본 논문에서 언급하고 있는 방법론 또한 이러한 방법론 중의 하나이다. 기존에는 핵심 문장을 추출하기 위하여 HTML의 경우 Title, 혹은 Index를 이용하거나, 혹은 신문과 같이 사건을 나열한 문서의 경우 문서의 전반부에서 핵심 문장을 추출하는 방법, 그 외에 가장 일반적으로 사용하는 통계 정보를 이용하는 방법이 있다[5-6]. 현재 가장 활발히 연구되고 있는 방법 중의 하나가 그 대상을 하나의 문서로 보고 이 문서를 구성하는 여러 개의 문단(Paragraph)들의 유사 관계나 중요도를 정보 검색 방법론을 적용하여 구분하고, 이를 토대로 요약문을 생성하기 위한 중요 문단을 추출하는 방식이다[4, 7-8]. 이는 문단간의 관계를 통계적 정보를 이용하여 구한 후 각 문단들 간의 유사도(similarity) 관계를 하나의 path로 보고 이를 따라가면서 중요한 문장을 선택하는 방식이다[4, 9].



(그림 1) 문단의 유사관계 그래표



(그림 2) 분할된 문단의 유사관계 그래프

(그림 1)은 문단의 유사도 관계 그래프를 보여준다. 각각의 노드는 문서의 각 문단과 해당 문단 번호를 나타낸다. 유사도 관계 그래프의 생성은 일반적으로 두 문단의 내용이 유사한지를 살펴보며 유사도가 0.2 이하인 링크는 생성하지 않는다. (그림 1)은 각 노드의 밀집도가 평균 밀도가 5.00이며, 노드간의 평균 유사도는 0.40이다. (그림 2)는 (그림 1)이 분할된 형태를 보여준다. 각 노드 사이의 거리가 5이상인 링크는 무시하며, 인접한 링크만을 취한 형태이다. (그림 2)와 같은 분할된 형태는 문서의 요약을 작성하기 위해 사용

된다. 이때 Path를 선택하는 방법에 따라 요약의 결과가 다르게 나타날 수 있는데, 크게 Bushy, Depth-First, Segmented Bushy Path의 세가지 방법이 제시되고 있다.

- **Bushy path**

그래프 상에서 링크가 가장 밀집된 문단(노드)을 추출한 후 추출된 노드와 연결되어진 문단들을 문단번호 순으로 추출한다.

- **Depth-first path**

우선 가장 링크가 밀집된 문단(노드)을 추출한 후, 그 문단과 유사도가 높은 문단을 번호 순으로 추출한다.

- **Segmented Bushy path**

Bushy path와 유사한 방식을 가지나, (그림 2)와 같이 각각의 분할된 영역에서 가장 밀집도가 높은 문단(노드)들을 선택한 후 각각의 노드에 연결된 문단들을 문단 번호 순으로 추출한다.

이러한 문단 단위의 핵심 문장 추출은 문서의 크기가 큰 경우 효율적일 수 있으나, 문서의 크기가 작은 단문 혹은 중문의 경우 그 성능이 그리 좋지 않을 뿐만 아니라, 문단(paragraph) 단위의 핵심 문장 추출은 문단에 포함된 핵심 문장뿐만 아니라 불필요한 문장도 함께 추출하는 단점이 있다[3].

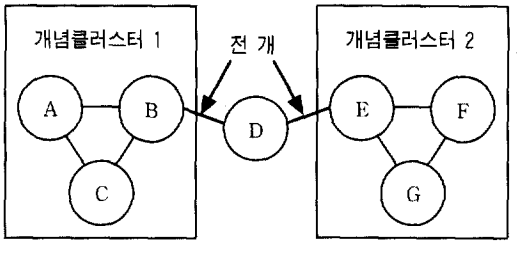
이러한 문단의 유사도 관계를 그래프로 표현하여 핵심 문장을 추출하는 방법은 최근의 연구 사례를 살펴보면 유사도 관계 그래프에서 이용된 Bushy, Depth-First, Segmented Bushy Path을 새롭게 확장한 경우를 살펴볼 수 있는데, 문단간의 병합을 통한 Traverse 방식을 이용하여 새로운 문단을 생성하고 이와 유사한 문단들을 추출해 나가는 방법을 사용하고 있다[3].

이밖에도 정보 검색 시스템에서의 질의확장을 이용하여 문서의 내용을 요약하는 방법이 제안되기도 하였는데, 간략히 살펴보면 요약하고자 하는 문서의 제목을 초기의 질의로 취하고, 초기 질의와 대상 문서내의 문장들간의 유사도를 계산하여 유사도가 높은 상위 몇개의 문장을 추출하여, 초기 질의와 병합하여 생성된 새로운 질의를 가지고, 다시 대상문서내의 문장들과의 유사도를 계산하여 유사도가 높은 문장들을 추출하는 방식이다[10-11]. 그러나 이 방식은 요약 문장들이 문서내에서 하나의 밀집된 클러스터를 이루고 있을 경우 추출된 핵심문장의 정확도가 낮아진다는 단점이 있다[1].

3. 공기 관계 그래프

본 논문에서는 기존의 핵심 문장 추출 방식이 가지는 단점을 극복하기 위하여 단어간의 관계를 이용하였다. 본 논문에서는 기존의 방식이 지니는 문단 위주의 핵심 문장 추출 방식이 아닌 대상 문서에서의 문장 및 문서를 구성하는 단어간의 공기 관계를 나타내는 그래프를 생성하고, 이를 이용하여 핵심어들을 추출한 후, 이를 기반으로 하여 핵심 문장을 추출하는 방식을 사용하였다.

이 방식은 문서가 저자의 독자적인 생각을 주장하기 위해 쓰여졌다는 가정하에 문서상의 저자의 주장을 대표하는 키워드 추출에 효과적인 방식이다[12]. 공기 관계 그래프의 구조는 문서를 크게 개념클러스터(Mean-Cluster), 주장(Insistence), 전개(Deployment)로 구분한다.



(그림 3) 공기 관계 그래프의 구성

공기 관계 그래프는 문서에서 출현 빈도가 높은 단어들의 공기 관계를 이용하여 구성한다. 문서에서 출현 빈도가 높은 단어들은 개념클러스터 및 키워드, 주장의 후보 단어가 된다. 출현 빈도가 높은 단어들은 문서의 요점과는 상관 관계가 적을 수는 있으나 저자가 주장을 펼치거나 혹은 설명을 하는데 있어서 당연시되는 전제들이다. 공기 관계 그래프에서 각각의 구성요소는 다음과 같이 정의된다.

- 개념클러스터

어떤 문장에서 같이 나타나는 단어들이 다른 문장에서 같이 나타난다면, 그 단어들은 그만큼 문서전체에서 중요성을 나타내는 개념들로 생각된다. 따라서 개

넘클러스터는 문서의 여러 문장에서 공통적으로 나타나는 단어들의 집합이라고 할 수 있다. 이 단어들의 집합들은 공기 그래프상에서 루프를 형성하게 된다. 그래프 상에서 루프를 형성한다는 것은 문서상에서 같은 문장 혹은 문단에서 동시에 나타나는 것을 의미하며, 어떠한 문장 혹은 문단이 자주 출현한다는 것은 저자가 주장하거나 혹은 설명하는데 있어서 주된 내용과 밀접한 연관을 가지고 있음을 의미한다.

● 주 장

저자가 의도하는 문서의 핵심이 될 수 있으며, 개념클러스터 사이에 강하게 연결되어 문장을 통합하는 역할을 가진다.

● 전 개

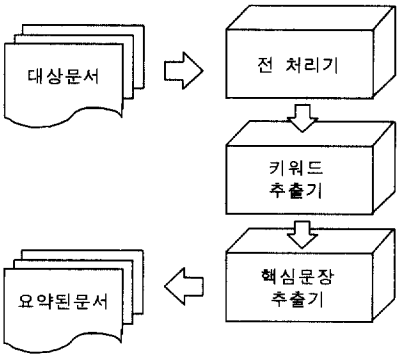
개념클러스터간의 연결을 나타내는 것으로서 문서에서 중요한 내용의 흐름을 표현한다. (그림 3)에서 B와 E는 문서의 키워드가 되는데 이는 문서에서 주장과 강하게 연결되어 저자의 의도를 뒷바침하거나 혹은 강조하는 역할을 하는 단어를 의미한다[13].

4. 핵심 문장 추출

본 논문에서는 저자의 글에서 나타내고자 하는 내용과 밀접한 관계를 가지는 핵심어들을 추출한 후 이 핵심어들을 이용하여 핵심 문장을 추출한다. 일반적으로 문서의 내용을 대표하는 핵심 문장은 저자가 의도하고자 하는 바와 밀접한 연관을 가지는 단어들(핵심어)을 한 개 이상 포함하고 있다고 볼 수 있으며, 핵심어를 많이 포함하는 문장일수록 문서에서의 중요도는 높다고 볼 수 있다. 이에 본 논문에서는 이러한 문장들을 핵심 문장으로 간주하였으며, 문장의 중요도를 결정하는 핵심어들을 추출하기 위하여 단어의 공기 관계 그래프를 이용한 키워드 추출 기법을 이용하여 핵심어들을 추출하고, 이를 이용하여 핵심 문장들을 추출하는 방법을 사용하였다.

4.1 핵심 문장 추출 시스템의 구성

본 논문에서 제안된 핵심 문장 추출기는 크게 전처리기, 키워드 추출기, 핵심 문장 추출기의 3개 모듈로 구성된다.



(그림 4) 핵심 문장 추출기 시스템 구성도

요약되어 질 문서들은 우선 전 처리기에서 후보 단어를 분류하게 된다. 한국어 문서의 경우는 형태소 분석 및 불용어 처리 과정을 거쳐 후보 단어를 분류하며, 영어의 경우에는 스태밍작업을 거쳐 후보 단어를 분류한다[14]. 본 시스템에서는 각각의 품사 중에서도 명사, 형용사, 동사만을 키워드의 후보 단어로 설정하며, 단어들에 대해서 중복을 제거하고, 기타 불필요한 요소를 제거한다[15]. 이렇게 각각의 준비 모듈을 통해 선정된 후보 단어에 대해서 키워드 추출기는 단어간의 공기 관계 그래프를 생성하고, 이를 이용하여 키워드를 추출한다. 추출된 키워드는 핵심 문장 추출기에서 해당 키워드들을 이용하여 핵심문장을 추출하게 되며, 이렇게 추출된 핵심 문장들이 정리되어져 문서의 내용을 요약하게 된다.

4.2 전 처리 단계

요약할 문서는 키워드를 추출하기 위해서 우선 형태소 분석 혹은 스태밍 작업을 우선 수행하며, 불용어(Stop List)등 불필요한 단어를 제거하고, 단어의 중복을 제거한다. 본 논문에서는 숙어의 요소는 고려하지 않았으며, 단어의 선정도 명사, 형용사, 동사로 제한하였다. 후보 단어의 집합은 다음과 같이 표현한다.

$$Document = W1, W2, W3, W4, \cdots\cdots Wn$$

4.3 키워드 추출

4.3.1 개념클러스터의 형성

개념클러스터(Mean-Cluster)를 형성하기 위해서 우선 각 단어들로 이루어진 공기 관계 그래프를 생성한

다. 그래프를 생성하기 위해 대상 문서에서 출현빈도의 순위별로 후보 단어 집합으로부터 상위 30위까지의 단어를 추출하여 그래프의 Vertex로 간주한다. 만약 30위에 해당하는 단어가 복수 개일 경우 해당 단어를 모두 선택한다. 그래프를 형성하는데 있어서 출현빈도가 높은 단어들을 추출하는 이유는, 저자가 문서에서 어떠한 주요 내용을 서술할 때는 중요한 내용과 밀접한 관련을 가지는 단어들이 반드시 그 문장내에 존재하기 마련이며, 이러한 단어들이 반드시 문서내에서 자주 나타난다고 볼 수는 없으나, 중요한 내용에 해당하는 단어 자체만으로 문장을 구성할 수는 없으므로 반드시 그와 함께 자주 나타나는 단어들은 존재하고, 따라서 문서내에서 자주 사용되는 단어일수록 그러한 중요한 단어와 같이 쓰일 확률이 높은 것으로 간주할 수 있다. 이는 자주 출현하는 단어들의 집합에서 추출된 개념클러스터내의 단어가 문서내에서 저자가 표현하려는 내용과 밀접한 관계를 가지는 단어와 함께 출현할 확률이 높은 단어들로 간주될 수도 있음을 나타낸다.

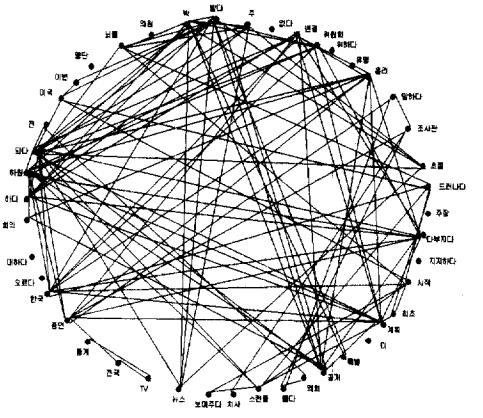
본 논문에서는 추출된 Vertex들의 공기관계를 나타내기 위해서 Vertex들의 쌍을 만들어 공기도를 계산하고, 계산된 공기도가 유효한 경우 즉 문서에서 한번 이상 동일한 문장에서 나타난 경우에만 Vertex간의 관계를 나타내는 Edge를 생성해준다. 공기도는 두 개의 단어가 대상 문서내의 동일한 문장에 대해 동시에 출현한 횟수의 합을 의미한다.

$$Co(Wi,Wj) = \sum_{S \subset D} |Wi|_S |Wj|_S$$

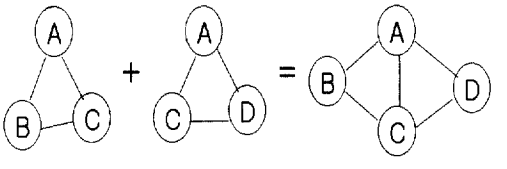
S : 대상 문서에 존재하는 문장
D : 대상 문서
$|W|_S$는 문장 S에서 단어 W의 출현 빈도

일단 그래프를 생성하면 그래프 상의 단어들의 연결관계를 분석하여 개념클러스터를 생성한다. 개념클러스터의 단어들은 문서에서 주장, 키워드의 후보단어가 된다. 각각의 개념클러스터는 문서에서 중요한 내용의 흐름을 나타내는 문장 혹은 문단에 속하는 단어들의 집합을 의미한다. 개념클러스터는 그래프 상에서 루프를 형성하는 단어들의 집합으로 표현되며, 개념클러스터를 찾는 과정은 (그림 6)과 같다.

개념클러스터를 찾기 위해서 우선 각각의 노드가 가지는 연결 노드 정보를 가지고 우선 노드 3개로 구성된 가장 간단한 Loop의 형태들을 찾는다. 이러한 각

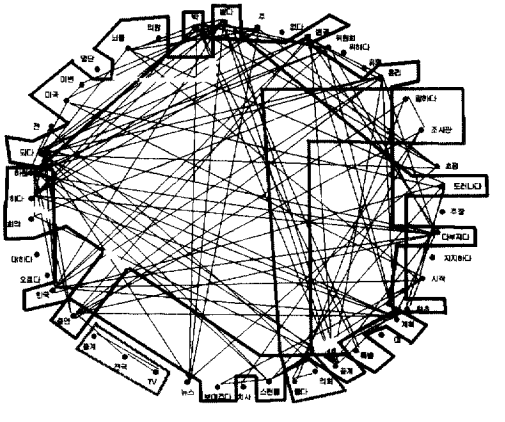


(그림 5) 공기관계 그래프



(그림 6) 그래프에서 루프를 찾는 과정

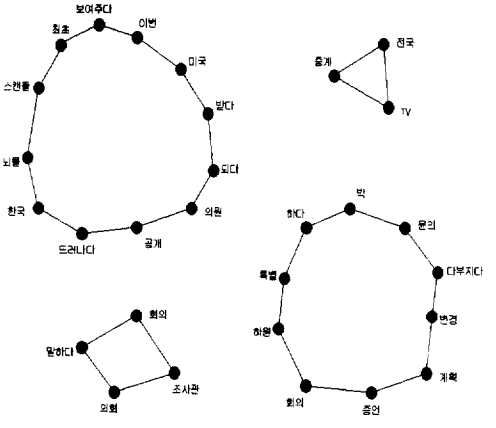
각의 Loop들은 서로 동일한 Edge를 가지는 Loop들과 다시 합쳐 새로운 형태의 Loop 그룹을 형성한다. 이러한 과정을 계속 반복하여 더 이상 동일한 Edge를 가지지 않는 Loop가 형성될 때까지 반복한다. (그림 7)은 이러한 과정의 최종 결과를 보여준다.



(그림 7) 형성된 개념클러스터

(그림 7)에서 굵은 선으로 표시된 부분 중 돌출된 부분에 해당하는 단어들의 집합이 하나의 개념클러스

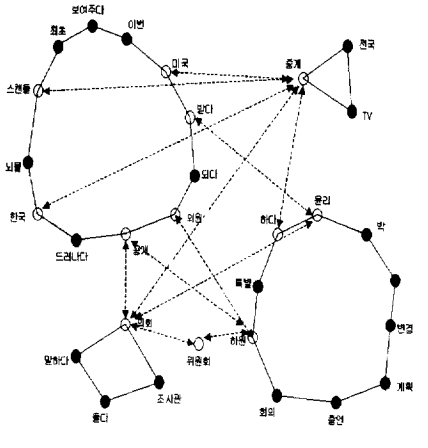
터를 형성한다[12]. (그림 7)에는 총 4개의 개념클러스터가 존재하며 각각의 개념 클러스터 별로 최적화한 결과는 (그림 8)과 같다.



(그림 8) 최적화된 개념클러스터의 집합

4.3.2 주장의 형성

주장은 문서내에서 저자가 의도하는 내용의 흐름을 나타내는 단어로서 개념클러스터와 개념클러스터를 강하게 연결시켜주는 역할을 한다. 주장은 각각의 개념클러스터상에서의 존재 여부에 관계없이 각각의 개념클러스터들과 강하게 연결된 단어를 나타낸다. 개념클러스터는 저자가 의도하는 중요한 내용의 일부로서 각각의 개념클러스터를 서로 연결 시켜준다는 것은 문서에서 중요한 내용들을 저자가 의도하는 내용상의 일괄된 흐름을 가지고 연결 시켜주는 것을 의미한다.



(그림 9) 주장과 전개의 추출

주장은 개념클러스터의 단어들과 강한 연결성 즉 높은 공기도를 가지는데 (그림 7)에서 링크의 밀집도가 높은 단어들이 이에 해당하며, 또한 두 개 이상의 개념클러스터내의 단어들에 대해서 동시 출현 확률이 높다. 주장(Insistence)을 계산하기 위한 함수 Key(W)는 임의의 단어 W가 각각의 개념클러스터에 속하는 단어들에 대해서 동시에 출현할 수 있는 확률을 계산하며, W가 특정 개념클러스터에 해당할 경우에도 자신이 속한 개념 클러스터에 대한 동시 출현 확률을 계산한다. 본 논문에서는 Key(W)값이 높은 순위부터 상위 12위까지의 단어를 주장으로서 선택하였다.

$$Key(W) = [1 - \prod_{c}^{cluster} (1 - f(w,c)/F(c))]$$

$$F(c) = \sum_{s \subset D} \sum_{w \subset D} |C - W|_s$$

$$f(w,c) = \sum_{s \subset D} |W|_s |C - W|_s$$

$$|C - W|_s = |C|_s - |W|_s \quad \text{if } W \subset C$$

$$= |C|_s \quad \text{if } W \not\subset C$$

$$|C|_s = \sum_{X \subset C} |X|_s$$

Clusters : 개념클러스터의 개수
f(W,C) : 단어 W와 개념클러스터 C와의 공기도
F(C) : 모든 단어와 모든 개념클러스터의 공기도의 합.
S : 문장 W : 단어 C : 개념클러스터
X : 개념클러스터 C에 포함되는 단어.
|X|s : 문장 S에 대한 개념클러스터 C에 포함되는 단어 X의 출현빈도

(그림 9)는 Key(W)에 의하여 계산된 각각의 그래프상의 노드로부터 주장과 전개를 추출한 결과를 보여주고 있다. (그림 9)에서 투명한 점으로 표시된 노드들이 주장에 해당하며, 각각의 점선 화살표들은 주장과 개념클러스터사이의 전개를 나타낸다.

4.3.3 키워드 추출

키워드는 문서에서 저자의 주장을 강하게 뒷바침하거나 혹은 자세한 설명을 위해 쓰여진 단어로서 그래프 상에서 주장과 강하게 연결되어지는 개념클러스터내의 단어를 의미한다. 즉 주장과 함께 자주 출현하는 단어는 그만큼 저자의 의도를 설명하기 위해 자주 사용되었다는 것을 의미하며 이는 키워드로 간주될 수 있다, 키워드는 그래프 상에서 주장과의 공기도가 높은 개념클러스터내의 단어를 의미한다. 단 어떠한 주장1이 다른 주장2와 높은 공기도를 가지며, 주장1과 주장2가 모두 특정 개념클러스터내의 단어일 경우 주

장1과 주장2는 우선적으로 키워드로서 간주된다.

키워드의 추출은 그래프를 이용해서 얻어진 주장 중의 단어 Wi와 개념클러스터에 포함되는 단어 Wj 사이의 강도를 계산하여 강도가 높은 단어를 키워드로 선택한다. 본 논문에서는 강도가 높은 순위부터 상위 12위까지의 단어를 키워드로 선택하였다. 주장과 개념클러스터의 강도 INTENSITY(K)는 개념클러스터에 있는 임의의 단어 K가 각각의 주장과 가지는 강도의 합을 나타내며, 단어 K의 가중치로 간주된다.

$$\text{Intensity}(K, INS) = \sum_{S \subset D} | K \| INS | S$$

$$\text{INTENSITY}(K) = \sum_{i}^{Insistence} Intensity(K, INSi)$$

KeyWord = min(12, | Document |)
K : 각각의 개념클러스터에 존재하는 키워드 후보 단어
INS : 주장.
Insistence : 주장의 개수

앞의 식에서 나타나듯이 키워드의 강도 INTENSITY (k)를 계산하는 과정에서 하나의 키워드 후보가 모든 주장과 가지는 공기도의 합을 키워드의 강도로서 취하는 이유는 하나의 개념클러스터에 대해 한 개 이상의 주장이 강하게 연결될 수 있기 때문이다. (그림 9)에서 주장과 각각의 개념클러스터사이의 강도를 계산하여 추출한 키워드의 목록 및 주장과의 강도는 5장의 <표 5>에 나타나 있다.

4.4 핵심 문장의 추출

추출된 키워드들은 핵심 문장 추출기의 후보 문장을 선정하는 기준이 된다. 핵심 문장은 우선 문장의 흐름을 대표할 수 있는 주장에 해당하는 단어를 포함하며, 주장과 강하게 연결되어지는 다른 키워드를 포함 할수록 해당 문서에서의 중요도가 높다. 즉 본 논문에서는 키워드를 포함하는 모든 문장을 우선 핵심 문장의 후보로서 추출한다

$$SK = sk_1 + sk_2 + sk_3 + sk_4 + \cdots\cdots sk_n (SK \subset D)$$

SK : 키워드를 포함하는 문장의 집합.
sk_n : 키워드를 포함하는 문장.
D : 대상 문서

일단 키워드를 포함하는 문장들을 추출하고 나면 추출된 문장들 중에서 주장을 포함하는 문장을 찾는다.

$$SI = si_1 + si_2 + si_3 + si_4 + \cdots\cdots si_n (SI \subset SK)$$

SI : 키워드를 포함하면서 주장도 같이 포함하는 문장의 집합
si_n : 키워드를 포함하면서 주장도 같이 포함하는 문장

추출된 키워드와 주장을 모두 가지는 문장들을 추출한 후에는, 추출된 문서각각에 포함된 키워드들이 가지는 가중치의 합이 높은 문장에 중요도를 높게 부여하는 방식으로 핵심 문장을 추출한다. 문장의 중요도를 구하는 함수 Importance(SI)는 다음과 같다.

$$\text{Importance}(SI) = \sum_{k \in SI} INTENSITY(k)$$

K : Si에 포함된 키워드의 집합

5. 실험 및 평가

문서의 요약에 대한 특별한 평가 기준은 아직까지 마련되어 있지 않다. 본 논문에서는 제안된 방법의 실험 및 평가를 위해서 사회 영역과 인문 영역에서 각각 40개의 문서를, 그리고 시사주간지인 타임지의 내용을 이용하여 구성된 한영 병렬 대역어 코퍼스 문서를 한글과 영어 각각 40개씩을 추출하여 실험 및 평가를 하였다. <표 1>부터 <표 6>은 대상 문서 중 시사 주간지인 타임지의 문서 중 한 개를 선택하여 단어의 공기 관계 그래프를 이용한 핵심 문장 추출 방식에 의한 추출 과정을 보여준다.

〈표 1〉 표본 문서 (단어 수 : 405)

이번주 박씨는 다시 뉴스의 초점을 받게 된다. 막바지에 계획이 변경되지 않는다면 다부진 이 한국인은 하원 윤리 위원회의 공개 회의에서 증언을 하게 된다. ….
……(중략)……
그의 증언은 1년 6개월 전부터 드러난 한국의 뇌물 스캔들을 최초로 완벽하게 미국 국민들에게 보여주게 된다.

〈표 2〉 상위 출현 빈도어 (30위까지)

단어	출현빈도	단어	출현빈도	단어	출현빈도	단어	출현빈도
받다	13	박	13	의원	9	특별	5
뇌물	5	명단	4	이번	4	미국	4
되다	4	하다	4	전	4	위원회	4
하원	4	대하다	3	오르다	3	한국	3
증언	3	윤리	3	회의	3	TV	3
치사	3	의회	3	변경	3	유명	2
주장	2	시작	2	공개	2	스캔들	2
중계	2	보여주다	2	전국	2	돌다	2
미	2	지지하다	2	최초	2	계획	2
초점	2	다부지다	2	조사관	2	말하다	2
뉴스	2	드러나다	2	위하다	2	없다	2
주	2						

〈표 3〉 문서에서 나타나는 개념클러스터

개념클러스터	단 어
1	이번, 미국, 받다, 되다, 의원 공개, 드러나다, 한국, 뇌물, 스캔들, 최초, 보여주다.
2	박, 하다, 특별, 하원, 회의, 증언, 계획, 변경, 다부지다, 윤리
3	들다, 의회, 말하다, 조사관
4	전국, TV, 중계

〈표 4〉 문서에서 나타나는 주장 (12위까지)

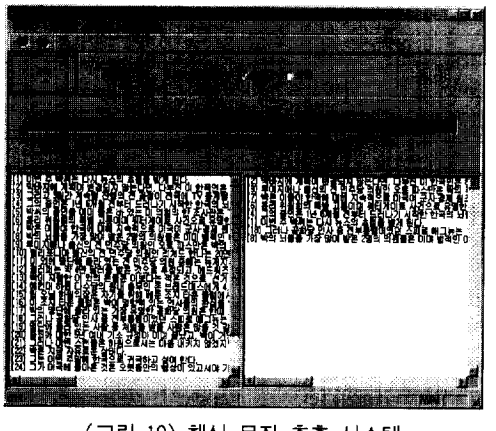
주장	연결된 개념클러스터	Key(W)	주장	연결된 개념클러스터	Key(W)
받다	1-2	0.1404	의원	1-2	0.1241
미국	1-3	0.1187	한국	1-4	0.1079
위원회	2-3	0.0917	윤리	2-3	0.0917
공개	1-3	0.0917	하다	2-3, 2-4	0.0864
스캔들	1-4	0.0863	전	3-4	0.0756
하원	2-1	0.0756	시작	2-4	0.0702

〈표 5〉 키워드 및 주장과의 강도 (12위까지)

키워드 및 해당 클러스터	주장과의 강도 및 주장	키워드 및 해당 클러스터	주장과의 강도 및 주장
받다(1)	33	되다(1)	19
박(2)	19	뇌물(1)	18
하다(2)	15	의회(3)	12
공개(1)	12	미국(1)	10
한국(1)	9	하원(2)	8
증언(2)	7	특별(2)	7

〈표 6〉 추출된 핵심문장

[1] 이번 주 박씨는 다시 뉴스의 초점을 받게 된다.
[2] 막바지에 계획이 변경되지 않는다면, 다부진 이 한국인은 하원 윤리 위원회의 공개 회의에서 증언을 하게 된다.
[4] 그의 증언은 1년 6개월 전부터 드러나기 시작한 한국의 뇌물 스캔들을 최초로 완벽하게 미국 국민들에게 보여주게 된다. … (후략)



(그림 10) 핵심 문장 추출 시스템

- 처리 속도(문서에 존재하는 후보 단어 개수를 기준으로)

  400단어 미만 : 25-30초/문서
  400단어 이상 1000단어 미만 : 30-120초/문서
  1000단어 이상 : 120초 이상/문서

- 테스트 환경(문서에 존재하는 후보 단어 개수를 기준으로)

  CPU : Pentium III 550MHz, 512cache
  RAM : 256M
  운영체제 : Windows2000 Professional, Windows-NT4.0 Workstation, Windows98

(그림 10)은 실제 구현된 시스템의 실행 화면이며 <표 1>부터 <표 6>까지의 과정과 동일한 문서에 대해서 실행된 모습을 보여주고 있다. 실험 결과의 평가는 두 사람이 동일한 문서에 대해서 추출한 핵심 문장과 본 논문에서 구현한 시스템의 결과를 비교하여 추출된 문장의 일치율을 평가 기준으로 하였다[4]. 평가를 위해 대학생 및 대학원생 7명이 참여하여 각각의 대상 문서에서 핵심 문장을 추출하였으며, 핵심 문장의 크기는 각 대상 문서의 30%로 고정하여 추출하였다. 각각의 수작업에 의한 결과는 서로 다를 수 있으며 그 결과의 평균 일치도(문장별)는 48.2%이다. 각각의 평가는 다음의 네 개의 기준으로 수행하였다.

- **Optimistic evaluation**

각각의 수작업에 의한 결과와 구현된 시스템의 결과를 분석하여 일치도가 높은 것을 선택한다.

- **Pessimistic evaluation**

각각의 수작업에 의한 결과와 구현된 시스템의 결과를 분석하여 일치도가 낮은 것을 선택 한다.

- **Union**

각각의 수작업에 의한 결과를 합쳐서 문장들을 추출하고 그 결과와 시스템의 결과를 비교한다.

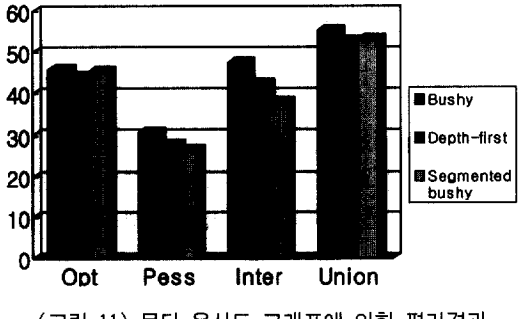
- **Intersection**

각각의 수작업에 의한 결과 중에서 일치하는 문장들을 추출하고 그 결과와 시스템의 결과를 비교한다.

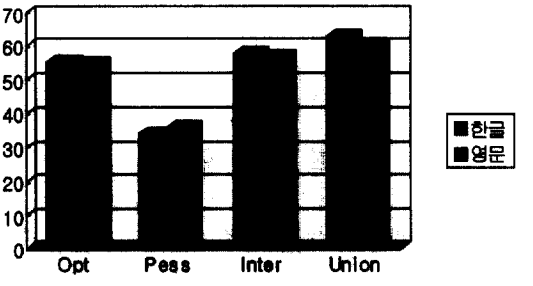
실험 결과에 있어서 수작업에 의한 결과와 본 논문

에서 구현된 시스템의 결과와의 일치도가 높을수록 시스템의 성능이 사람에 의한 수작업에 근접한 수준으로 간주한다. 특히, 본 논문에서는 앞의 네 가지 정의된 평가 방법 중에서 Intersection의 결과에 가장 초점을 맞추고 있다. 동일한 문서에 대하여 두 명 이상의 사람이 핵심 문장을 추출한 결과들의 일치도가 낮다는 것은 두 사람의 교육 수준이 비슷하면서도 문서의 내용을 파악하는 관점이 상당 부분 일치하지 않는다는 것을 의미하지만, 서로 다른 견해를 가지는 두 명 이상의 사람이 동일한 문장을 핵심 문장으로 선택했다는 것은 그만큼 일치된 추출 문장들은 문서의 내용면에서 중요한 부분을 포함하고 있다고 간주될 수 있기 때문에 본 논문에서는 구현된 시스템에 의해 추출된 핵심 문장들이 두 명 이상이 수행한 수작업의 일치 부분과 얼마나 많이 일치하는지에 그 결과의 초점을 맞추고 있다.

본 논문에서는 시스템 성능의 객관적 평가를 위해서 우선 관련연구에서 언급한 문단의 유사도 관계 그래프를 이용한 방법을 동일한 문서에 대해서 Bushy, Depth-first, Segmented bushy path의 방법으로 각각 추출한 핵심 문장들과, 역시 동일한 대상 문서에 대해서 수작업에 의한 추출결과와의 평가결과(그림 11)를 본 논문에서 제시한 단어의 공기 관계 그래프를 이용한 핵심 문장 추출 방법에 의한 평가결과(그림 12)와 비교함으로써 성능평가를 수행하였다. (그림 11)는 기존의 문단의 유사도 그래프를 이용하여 핵심문장을 추출한 결과와 수작업에 의한 결과의 평균 일치도를 나타내며 (그림 12) 역시 본 논문에서 구현된 단어의 공기 관계 그래프를 이용한 핵심 문장 추출 결과와 수작업에 의한 결과의 평균 일치도를 나타내고 있다.



(그림 11) 문단 유사도 그래프에 의한 평가결과



(그림 12) 공기 관계 그래프의 평가결과

6. 결 론

본 논문에서는 단어의 공기 관계 그래프를 이용한 핵심 문장 추출 방법을 제안하고 있다. 실험 결과 본 논문에서 제안한 핵심 문장 추출 방법은 기존의 방식보다 그 일치율이 평균 10% 정도 높은 것으로 평가되었다. 특히 주목할 점은 두 사람 공통으로 추출한 문장이 문서내에서 중요한 문장일 확률이 높다고 볼 때 추출된 문장들과 본 논문에서 제안한 방법으로 추출한 문장간의 일치율을 비교한 결과(Intersection)가 10%이상 향상되었다는 것이다. 본 논문에서 제안한 핵심 문장 추출 방법의 특성상 그 핵심 문장 추출 결과의 정확도는 키워드 추출의 정확도에 의해서 많은 영향을 받는다. 본 논문에서 이용된 키워드 추출기는 키워드의 후보로서 기존의 키워드 추출기와는 달리 명사 외에도 동사, 형용사 등을 키워드로 함께 추출하기 때문에 그 정확도를 기존의 키워드 추출 방법과 비교 평가하기 어려운 점이 있으나, 문서의 내용을 요약하기에는 기존의 키워드 추출 방법을 사용하는 것보다는 그 정확도가 우수한 것으로 평가된다. 또한 영어 문서에 대해서 핵심 문장을 추출한 결과들도 한국어 문서와 거의 유사한 수준의 정확도를 보이고 있으며, 이는 한국어만이 아닌 다른 언어에 대해서도 문서의 핵심 문장 추출 결과에 대한 기대치를 높게 가질 수 있는 것으로 평가된다.

다만, 실험 결과에서 알 수 있듯이 기존에 연구되어진 단락의 유사도 관계를 이용하여 추출한 방법보다 우수한 성능을 보이고는 있으나, 문서의 길이가 장문에 해당하는 경우 그 수행 속도가 현저히 떨어지는 단점을 가지고 있으며, 문서의 크기가 커질수록 후보 단

어의 개수 선정에 대한 어려움 및 후보 단어의 수가 늘어날수록 지수적으로 증가하는 수행 속도의 개선 문제 등이 앞으로 해결할 과제로 남아 있다.

참 고 문 헌

[1] 한경수, 백대호, 임해창, "질의 확장을 이용한 자동 문서 요약", 한국정보과학회 학술논문발표지 제27권 1호 pp. 339-341, 2000.

[2] Julian Kupiec, Jan Pedersen, and Francine Chen, "A Trainable Document Summarizer," Proceedings of ACM-SIGIR'95, pp.68-73, 1995.

[3] 유동원, 이종혁, "단어 공기 정보를 이용한 자동화 문서요약", 한국정보과학회 학술논문발표지 제27권 1호 pp. 345-347, 2000.

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

[5] H. P. Edmundson, "New Method in Automatic Extracting," Advances in Automatic Text Summarization, pp.23-42, MIT Press, 1999.

[6] Julian Kupiec, Jan Pedersen, Francine Chen, "A Trainable Document Summarizer," Proceedings of ACM-SIGIR'95, pp68-73, 1995.

[7] G.Salton, A.Signal, "Automatic Theme Generation and the Analysis of Text Structure," TR, 1994.

[8] Regina Barzilay, Michael Elhadad, "Using Lexical Chains for Text Summarization," Advances in Automatic Text Summarization, pp111-121, MIT Press, 1999.

[9] G. Salton, A. Signal, C. Buckley, M. Mitra, "Automatic Text Decomposition Using Text Segments and Text Theme," 96 ACM Conference on Hypertext, 1996.

[10] Edward Hovy, Chin-Yew Lin, "Automated Text Summarization in SUMMARIST," Advances in Automatic Text Summarization, pp.81-94, MIT Press, 1999.

[11] Daniel Marcu, "Discourse trees are good indicators of importance in text, Advances in Automatic Text Summarization," pp.123-136, MIT Press, 1999.

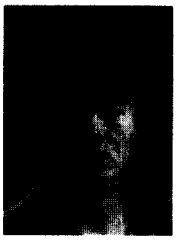
[12] Yukio Ohsawa, Nels E. Benson, and Masahiko Yachida, "Automatic Indexing by Segmentation and Unifing Co-occurrence Graphs," 전자정보통신학회 논문지 D-1 Vol.J82-D-I, No.2, pp391-400, 1992.

[13] 류 제, 한광록, "단어의 공기 관계 그래프를 이용한 인터넷 문서의 키워드 추출", HCI2000 학술대회발표 논문집 9권 1호, pp.894-899, 2000.

[14] 서영훈, 이하규 외, "한국어 구문 Tagged Corpus 구축 및 구문 분석 데이터 사전 개발", 한국 전자통신 연구소 최종 연구 보고서, 1998.

[15] 강승식, "음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석", 서울대학교 대학원 컴퓨터공학과 박사 학위 논문, 1993.

[16] Ellis Horowitz, Sartaj Sahni, Dinesh Mehta, "Fundamental of data structures," p330-396.



류 제

e-mail : ryuje@shinbiro.com

1999년 호서대학교 컴퓨터공학과 졸업(학사)

1999년 호서대학교 컴퓨터공학과 석사과정 입학

2000년 호서대학교 벤처 전문대학원 편입

관심분야 : 에이전트, Mobile 컴퓨팅, 자연어 처리.

한 광 록

e-mail : krhan@office.hoseo.ac.kr

1984년 인하대학교 전자공학과 졸업(공학사)

1986년 인하대학교 대학원 정보공학전공(공학석사)

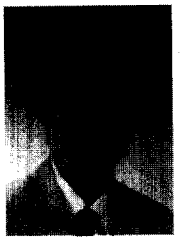
1989년 인하대학교 대학원 정보공학전공(공학박사)

1989년~1991년 한국체육과학원 선임 연구원

1991년~2000년 현재 호서대학교 컴퓨터공학부 교수

1999년~현재 호서대학교 벤처전문대학원 컴퓨터응용기술팀장

관심분야 : 정보검색, 자연언어처리, 기계번역, HCI, 지능형 에이전트 등



손 석 원

e-mail : sohn@office.hoseo.ac.kr

1985년 인하대학교 전자 공학과 졸업(학사)

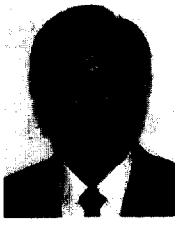
1987년 인하대학교 대학원 전자 공학과 졸업(공학석사)

1987~1992년 한국 원자력 연구소 선임 연구원

1995년~1997년 DX Net Internet Service 근무

1999년~현재 호서대학교 컴퓨터 공학과 전임강사

관심분야 : Radio Resource Management, CDMA power control, Internet User Interface, Internet Protocol



임 기 욱

e-mail : rim@omega.sunmoon.ac.kr

1977년 인하대학교 공과대학 전자 공학과 졸업

1987년 한양대학교 전자계산학 석사

1994년 인하대학교 전자계산학 박사

1977년~1983 한국전자기술연구소 선임연구원

1983년~1988년 한국전자통신연구소 시스템소프트웨어 연구실장

1988년~1989년 미 캘리포니아 주립대학(Irvine) 방문 연구원

1989년~1997년 한국전자통신연구원 시스템연구부장 주전산기(타이컴) III, IV 개발 사업책임자

1997년~2000년 정보통신연구진흥원 정보기술전문위원

2000년~현재 선문대학교 교수

관심분야 : 실시간 데이터베이스시스템, 운영체제, 시스템구조 등