동적 연결 그래프를 이용한 자동 문서 요약 시스템
(A Document Summarization System Using Dynamic Connection Graph)

송 원 문 † 김 영 진 † 김 은 주 † 김 명 원 ††
(WonMoon Song) (YoungJin Kim) (EunJu Kim) (MyungWon Kim)

요 약 문서 요약은 쉽고 빠르게 문서의 내용을 파악할 수 있도록 방대한 내용을 가지는 다양한 형태의 문서로부터 핵심 내용만을 추출하거나 생성하여 제공하는 것을 목적으로 한다. 본 논문에서는 효율적 문서 요약을 위해 주어진 문서의 평균 문장 길이(핵심어 개수)를 고려하여 문장 간의 핵심어 유사도를 나타내는 연결 그래프를 생성하고 분석하여 요약을 생성하는 기법을 제안한다. 또한 이러한 기법을 이용하여 응용 프로그램 문서로부터 자동으로 요약을 생성하는 자동 문서 요약 시스템을 개발한다. 제안한 방법의 객관적인 요약 성능 측정을 위해 정확한 요약문이 실린 20개의 테스트 문서를 이용하여 생성된 요약에 대해 precision(정확률)과 recall(재현율), F-measure를 측정하였으며, 실험 결과를 통해 기존 기법에 비해 우수한 요약 성능을 보임을 증명하였다.

키워드 : 문서요약, 동적 연결 그래프, 추출요약, 중요문장 추출, 핵심어 유사도

Abstract The purpose of document summarization is to provide easy and quick understanding of documents by extracting summarized information from the documents produced by various application programs. In this paper, we propose a document summarization method that creates and analyzes a connection graph representing the similarity of keyword lists of sentences in a document taking into account the mean length(the number of keywords) of sentences of the document. We implemented a system that automatically generate a summary from a document using the proposed method. To evaluate the performance of the method, we used a set of 20 documents associated with their correct summaries and measured the precision, the recall and the F-measure. The experiment results show that the proposed method is more efficient compared with the existing methods.

Key words : Document Summarization, Dynamic Connection Graph, Extractive Summarization, Keysentences Extraction, Keywords Similarity

1. 서 론

인터넷 기술의 발전으로 사용자들은 쉽고 빠르게 다양하고 방대한 정보를 접할 수 있게 되었다. 그러나, 가히 정보의 홍수라 할 만큼 너무 많은 정보가 무분별하게 제공되어 사용자들은 원하는 정보를 찾기 위해 더욱 많은 시간을 투자하여 제공되는 정보들을 일일이 확인해야만 한다. 이러한 문제는 정보 검색 뿐만 아니라 보안과 같은 분야에서도 동일하게 발생 한다. 최근 기업 내부자에 의해 파일 복사 등의 방법을 이용하여 기업의 기밀이 유출되는 사례가 발생하자, 기업들은 보안을 위하여 유출 행동을 로그로 기록하고 유출 내용을 파악하기 위해 원본 문서를 저장하는 솔루션들을 도입하고 있다. 그러나, 유출된 문서의 원본을 일일이 저장하고 내용을 확인하여 기밀문서 인지를 파악하기 위해서는 방대한 저장 공간과 함께 내용 확인을 위한 많은 시간을

· 본 연구는 숭실대학교 교내연구비 지원에 의해 이루어졌습니다.
· 이 논문은 2008 한국컴퓨터종합학술대회에서 '동적 연결 그래프를 이용한 자동 문서 요약 시스템'의 제목으로 발표된 논문을 확장한 것임

† 학생회원 : 숭실대학교 컴퓨터학부
gtangel@ssu.ac.kr
liebulia@ssu.ac.kr
blue7786@ssu.ac.kr
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
mkim@ssu.ac.kr

논문접수 : 2008년 8월 25일
심사완료 : 2008년 11월 7일

투자해야만 한다. 이와 같은 문제를 해결하기 위해 최근 정보 검색, 텍스트 마이닝 등의 기법을 이용한 문서 요약 연구들이 진행되고 있다.

문서 요약이란, 보다 적은 용량의 문서로 원본 문서에서 말하고자 하는 내용을 전달할 수 있도록 원본 문서로부터 가장 의미 있는 내용만을 파악하여 요약으로 구성하는 것을 말한다[1]. 이와 같은 문서 요약은 그 내용 구성 방법에 따라 생성요약과 추출요약으로 구분할 수 있다[2]. 생성요약은 원 문서로부터 중요한 단어들을 선별한 후 자연어처리 기법을 이용하여 새로운 문장을 구성하여 요약문으로 제공하는 것이며, 추출요약은 원 문서로부터 중요하다고 생각되는 문장만을 선별하여 요약문으로 제공하는 것이다. 전자가 인간 전문가에 의한 요약 생성과 유사한 방법이나 현재의 자연어처리 기술 한계와 제한점 때문에, 상대적으로 접근과 구현이 쉬운 추출 요약에 관한 연구가 주로 이루어지고 있다.

추출요약에서는 요약을 위한 각 문장의 중요도 계산과 함께 중복된 의미를 가지는 문장을 얼마나 효율적으로 제거하며 중요 문장을 추출하는가가 중요한 문제이다[3,4]. 이와 같은 문제의 해결을 위해 문장이 포함한 단어나 문장의 위치에 기반한 다양한 문장간 유사도 계산 기법들이 제안되었다[2,3,5]. 그러나 기존의 방법들은 같은 수의 공통 단어를 포함한 문장들이라도 개개의 문장이 포함한 단어 수나 문장의 출현 위치에 따라 유사도가 다르게 계산되는 단점이 있다.

본 논문에서는 이러한 단점을 해결하며 개체의 상관관계 표현을 위한 가장 적합한 구조인 연결 그래프 구조를 이용하여 원 문서를 문장간 연결 그래프로 표현하고, 표현된 그래프의 분할을 통해 중요 문장을 추출하는 방법[6]을 이용하되, 실생활의 다양한 형태의 문서에 적합한 요약 기법을 위해 문장간 연결 그래프 생성시 각 문장의 길이에 따라 연결 여부를 동적으로 결정하는 방법을 제안한다. 또한, 기존의 텍스트 추출 및 형태소 분석 방법을 통합하여, 응용 프로그램 문서로부터 자동으로 요약을 생성하는 시스템을 개발한다.

본 논문의 2장에서는 문서 요약에 관한 기존의 방법들을 간단히 알아보고, 3장에서는 제안하는 방법인 문장 길이에 따른 동적 연결 그래프 생성 방법과 자동 요약 시스템에 대하여 설명한다. 4장에서는 동적 연결 그래프를 이용한 요약의 평가를 위한 실험 결과를 기술하고 5장에서는 결론을 맺고 향후 연구 과제를 검토 한다.

2. 관련 연구

2.1 문장간 유사도 계산을 이용한 문서 요약

Cosine 유사도와 같은 계산을 이용하여 문서내의 문장들을 몇 개의 주제로 그룹화 하고 그룹별로 중요 문장을 추출하여 요약문으로 제공하는 방법은 단순한 계산방법을 사용함에도 불구하고 비교적 좋은 성능을 얻을 수 있으며, 중복 의미 문장 제거를 위한 문장 그룹화에 쉽게 응용할 수 있어 문서 요약을 위해 가장 많이 활용하는 방법이다.

최근의 연구로, [2]에서는 문서내의 각 문장을 문장의 위치 및 길이, 제목과의 공통 단어 포함 여부, 문장내 단어들의 항목 빈도 합, 고유명사 및 대명사 포함 여부 등으로 이루어진 벡터로 표현한 후 각 벡터간의 cosine 유사도를 계산하였다. 또한 [5]에서는 특정 그룹에 속한 각각의 문장들과 새로운 문장을 대상으로 공통으로 포함한 단어에 기반한 cosine 유사도를 계산하여 새로운 문장이 대상 그룹에 속할지의 여부를 결정하는 방법을 제안하였다.

이와 같이 수학적 방법을 통해 문장간 유사도를 계산하는 방법들은 기본적으로 문장을 특정 형태의 벡터로 표현해야만 한다. 그러나 유사도 계산을 위한 벡터 표현에 있어 단순히 출현한 단어나 문장의 길이, 위치 등의 정보만을 반영함으로써, 문장의 의미적 정보를 제대로 반영하지 못하고 문장이 포함한 단어 수나 길이에 민감하게 유사도가 계산된다는 단점이 있다.

2.2 그래프 분할을 이용한 문장 그룹 기반 문서 요약

[6]에서는 기존의 유사도에 기반한 문장 그룹화 방법들의 문제점을 해결하고자 그래프 구조를 이용하는 방법을 제안하였다. 이 방법에서는 먼저, 문장간에 공통 단어를 포함하고 있는지의 여부에 따라 문장의 연결여부를 설정하여 문서를 문장간 연결 그래프로 표현한다. 이후, 그래프의 이음새인 관절점(articulation point)이 되는 그래프의 노드(문장)를 찾아 그래프를 몇 개의 그룹으로 분할한 후 그룹별로 중요 문장을 추출하고 요약으로 제공한다. 특히, 문장간 연결을 위한 공통 단어 포함 여부판단을 위해서는 단어들간의 공기정보를 이용하여 추출된 키워드[7]를 이용함으로써 문장의 의미를 반영한 연결 구조 생성을 시도하였다.

이 방법은 문장의 길이나 위치에 의존적인 수학적 계산 없이 문장의 의미에 기반한 연결 구조만를 이용하여 좋은 요약 성능을 보였다는데 의의가 있다. 그러나, 단지 문장간에 공통으로 포함하는 단어가 있는지의 여부만을 판단하여 연결 그래프를 생성함으로써, 비교적 문장의 길이가 길고 내용이 자세히 기술되어 있는 기업의 기술 문서나 학술 정보와 같은 경우 문서 전체가 아주 적은 수의 그래프로 분할(최악의 경우에는 하나의 그래프만을 구성)될 가능성이 매우 높아 효율적인 요약을 기대할 수 없다.

2.3 문서 요약을 위한 문장 중요도 계산(중요 문장 추출) 방법

효율적인 문서 요약을 위해서는 앞서 언급한 연구들의 관점에서와 같이 문서가 포함하는 몇 개의 하위 주제별로 중요한 문장을 추출하기 위해 문장들을 비슷한 주제로 나누거나 구성하는 방법도 중요 하지만, 구성된 주제별 그룹에서 어떻게 중요 문장만을 선별하는가 또한 중요한 문제이다. 이와 같은 문제를 위한 최근의 연구들은 문장이 포함하는 중요 핵심어에 따라 중요 문장을 선별하는 방법[3,8,9]과 기계학습이나 수학적, 통계학적 기법을 사용하여 문장간의 상관관계를 분석하여 중요 문장을 선별하는 방법[2,10]의 두 가지로 나눌 수 있다.

문장이 포함하는 핵심어를 이용하여 중요 문장을 선별하는 방법은 비교적 직관적이면서도 그 성능이 다른 방법에 비해 높아 가장 많이 사용하는 방법이다. [3]에서는 핵심어로 선정된 단어가 많은 문장일수록 중요도가 높다고 판단하여 중요 문장을 추출하는 방법을 제안하였으며, [8]에서는 기 구축된 단어의 의미 사전을 이용하여 비슷한 의미의 단어가 많이 쓰인 문장을 중요 문장으로 추출하는 방법을 제안하였다. 이와 같은 방법들은 단순히 핵심어 개수만을 확인함으로써 긴 문장이 중요 문장으로 추출되는 부작용이 발생할 가능성이 높으며, 단어의 의미 사전이 미리 구축되어 있지 않을 경우 적용하기 힘들다는 단점이 있다. [9]에서는 이와 같은 문제를 해결하기 위해 핵심어의 중요도를 별도로 계산하여 문장 중요도에 반영함과 동시에 의미 사전 구축 없이 문장내의 단어들간의 의미적 관계나 구문적 관계를 판단하여 중요 문장을 선별하는 방법으로 핵심어들간의 공기 정보를 이용하는 방법을 제안하였다.

기계학습이나 수학적, 통계학적 기법을 사용한 방법은 문장들간의 상관관계를 분석하여 주변 문장에 더 많은 영향을 주는 문장을 중요 문장으로 선별하는 방법이다. [2]에서는 각 문장을 노드로 하여 홉필드 네트웍을 구성하고 학습하여 더 많은 주변 문장과 큰 가중치로 연결된 문장일수록 중요도가 높다고 판단하여 중요 문장을 선별하는 방법을 제안하였으며, [10]에서는 주성분 분석을 이용하여 문서내의 주성분이 되는 단어를 선별하고 문장과 주성분 단어를 이용하여 비정칙치 분해를 시행, 주성분 단어와 거리가 가깝다고 판단되는 문장들만을 선별하여 중요 문장으로 결정하는 방법을 제안하였다. 이와 같은 방법들은 중요 문장 판단을 위한 학습 시간이 오래 걸리거나, 그 구조가 복잡하여 구현이 어렵고 결과에 대해 이해하기 어렵다는 단점이 있다.

3. 동적 문장 연결 그래프를 이용한 문서 요약

본 연구에서는 기존에 제안된 가장 최근의 방법 중 단어의 공기정보에 따른 문장 연결 그래프를 이용한 문서 요약 방법에 대해 분석하고 이를 개선하여 보다 정확한 요약을 할 수 있는 방법을 제안하고자 한다. 특히, [6]에서 제안한 그래프 연결구조와 그룹 분할 방법을 이용하되 문장 연결 여부 판단을 동적으로 판단함으로써 보다 다양한 종류의 문서에 적용할 수 있는 효과적인 문서 요약 시스템을 제안한다.

3.1 문장 길이에 따른 동적 연결 그래프 생성

다음 그림 1의 경우, 문장간 연결 구조를 결정하는데 있어, 단지 공통단어를 포함하고 있는지의 여부만 확인한 것으로, 모든 문장이 순환 연결되어 표현된다(그림에서 내용을 표시하지 않은 연결은 두 문장간의 공통 단어가 5개 이상이라고 가정한다).

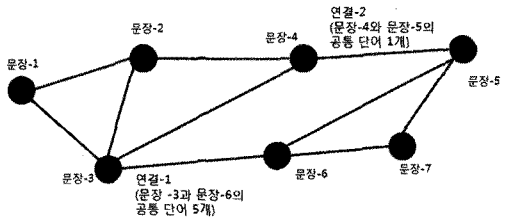


그림 1 공통 포함 단어를 이용한 문장 연결 그래프

[6]에서 제안한 이 방법에 의해 문장 연결 그래프를 생성하면, 원래의 문장 자체가 자세하고 길게 작성되어 많은 단어를 포함하고 있는 경우, 그만큼 다른 문장과 연결될 가능성이 많아지므로 문서가 순환 연결 그래프로 구성될 가능성이 높아진다. 본 논문에서는 불필요한 문장의 연결을 제한하여 순환 연결 문제를 해결하기 위해, 두 문장간의 공통 포함 단어의 수를 제한하여 문장의 연결 여부를 동적으로 결정하는 방법을 제안한다. 예를 들어 그림 1의 문장 연결 결정을 위한 문장간 공통 포함 단어의 수를 5개 이상으로 제한하면, '연결-2'는 사라지게 되며, 이로 인해 '문장-3'과 '문장-6'은 그래프를 분리할 수 있는 관절점 노드가 된다. 결과적으로 '연결-1'을 절단하여 문장들을 두 개의 그룹으로 분류하고 그룹별로 중요 문장을 추출함으로써, 중복문장이 제거된 효율적 요약을 생성할 수 있다.

문장 연결의 동적 결정을 위한 두 문장간 공통 포함 단어의 수의 제한 값은 문장의 길이에 따라 달라져야 한다. 이를 위해 본 논문에서는 먼저 [6]에서 사용한 공기정보를 이용한 키워드 추출 방법을 이용하여, 문서내의 주요 단어들을 추출하고, 각 문장을 주요 단어들의 리스트로 표현하였다. 이 후 다음 식 (1)과 같이 문장의 평균 포함 단어 수에 따라 제한 값을 결정하고 문장의 연결 여부를 판단한다.

$$\theta_{Connection} = \frac{Average(|Word_{Sentence}|)}{n}, \qquad (1)$$

$$CoOccur_{A,B} = |Words_{Sentence_A} \cap Words_{Sentence_B}|$$

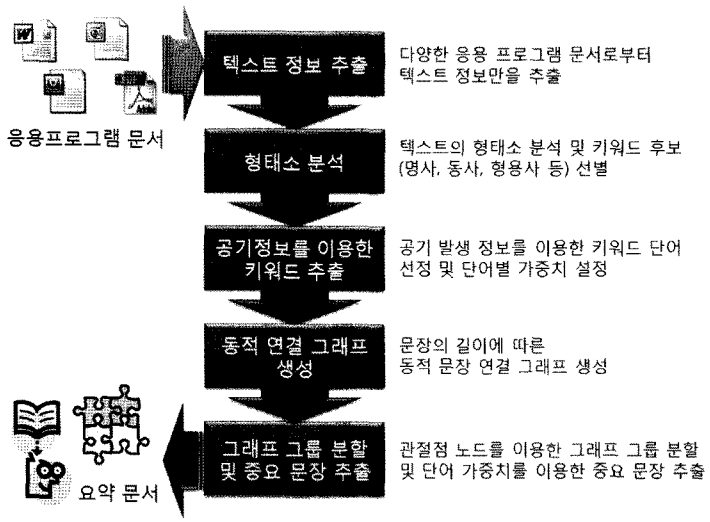


그림 2 자동 문서 요약을 위한 프로세스

$If\ CoOccur_{A,B} \geq \theta_{Connection}$,
$then\ Connect(Sentence_A, Sentence_B)$

$Average(|Words_{Sentence}|)$는 문서 내에 포함된 모든 문장의 평균 단어 수를 의미하며, n은 문장의 길이에 따라 $\theta_{Connection}$(공통 포함 단어 수의 제한 값)을 적절한 비율로 설정하기 위한 값으로 본 연구에서는 실험을 통하여 2로 정하였으며 이에 재한 실험 결과는 4.2절에 간단히 기술하였다. $|Words_{Sentence_A} \cap Words_{Sentence_B}|$는 임의의 두 문장 A와 B사이에 공통으로 포함된 중복되지 않은 단어의 수를 의미한다. 따라서, 이 값이 앞서 결정된 공통 포함 단어 수의 제한 값($\theta_{Connection}$) 이상인 경우에만 문장 A와 B는 그래프에서 연결된 것으로 표현한다.

3.2 자동 문서 요약 시스템

실생활에 문서 요약을 효과적으로 응용하기 위해서는 다양한 응용 프로그램 문서로부터 요약을 생성하여 사용자에게 제공하는 전 과정이 자동화된 문서 요약 시스템이 필요하나 현재까지의 문서 요약 연구들은 텍스트 문서나 단어의 조합으로 표현된 문장 구조를 가정한 중요 문장 추출 정도에 그치고 있다. 따라서 본 논문에서는 그림 2와 같이 제안한 방법과 기존에 개발된 텍스트 추출 및 형태소 분석 프로그램과 그래프 분할 기법을 융합하여 다양한 응용 프로그램 문서로부터 자동으로 요약을 생성하는 자동화된 문서 요약 시스템을 구현하였다.

자동화된 문서 요약을 위해서는 가장 먼저 다양한 응용 프로그램 파일로부터 텍스트 정보만을 추출하며 이를 위해 관련 기술 보유 업체[11]의 텍스트 추출 제공 라이브러리를 제공받아 이용하였다. 이렇게 추출된 텍스트 정보로부터 형태소 분석기를 통해 명사나 동사, 형용사 같은 의미 있는 품사의 단어만을 따로 추출한다. 형태소 분석에는 연구용으로 공개되어 있는 국민대의 KLT[12]를 이용하였다. 형태소 분석을 통해 키워드의 후보가 되는 의미 있는 품사의 단어들이 추출되면 추출된 단어들의 공기 정보 분석[6]을 통하여 문서의 주요 키워드 단어를 추출하고 가중치를 설정한다. 이렇게 추출된 키워드 단어들을 이용하여 문서내의 각 문장을 키워드 단어 벡터로 구성하고 제안한 방법을 통하여 동적으로 문장 연결 그래프를 생성한다. 생성된 그래프는 관절점을 이용한 그래프 분할 방법[6]을 통하여 몇 개의 그룹으로 나눈 후, 각 그룹별로 소속된 문장이 포함한 키워드 단어들의 가중치를 합산하여 그 값이 가장 높은 한 개의 문장을 그룹의 중요 문장으로 추출한다. 이때 i번째 키워드 단어인 T_i의 가중치($asim(T_i)$)는 다음 식 (2)와 같이 [6]에서 제안한 공기정보를 이용한 도함유사도를 이용하여 계산한다. 식 (2)에서 서 $Freq(T_i, T_j)$는 i번째 키워드 단어 T_i와 j번째 키워드 단어 T_j가 문서 내에서 동시에 출한 문장의 수이다. 그룹별로 추출된 중요 문장들을 문서내의 순서에 따라 정렬하면 사용자에게 제공할 요약 정보 생성이 완료 된다.

$$wsim(T_i, T_j) = \log(Freq(T_i, T_j)) \quad (2)$$
$$asim(T_i) = \sum_{j=1}^{n} wsim(T_i, T_j)$$

4. 실험 및 평가

제안한 요약 시스템의 성능 평가를 위한 테스트 문서

로는 IT 분야의 국내 전문 학술단체인 정보과학회, 인터넷정보학회, 정보보호학회 등에서 2000년 이후 발행한 학술지 논문 중 DBpia[13] 서비스를 이용하여 임의로 20건을 수집하였다. 이와 같은 문서는 문서 내에 본문과 함께 요약이 포함되어 있어 시스템의 성능 평가에 용이하다. 특히, 학술지 논문은 전문 심사관들에 의한 여러 차례의 심사과정을 통해 출판되는 문서들로써 비교 평가를 위한 문서내의 요약이 비교적 객관적이고 정확하게 작성되어 있다고 할 수 있다. 실험에 사용한 문서의 목록은 논문 마지막 페이지의 별지 1과 같다. 본 논문의 목적에 따라 신문기사, 웹 정보 등 다양한 형태의 문서를 수집하여 실험을 수행하여야 하나, 객관적으로 인증된 요약을 포함한 문서는 현실적으로 수집하기 어려우므로 실험은 수집한 학술지 논문으로 제한한다.

4.1 평가 방법

문서 요약의 평가방법으로는 ISI와 USC에서 공동으로 개발한 ROUGE(Recall-Oriented Understudy for Gisting Evaluation)[14] 방법이 가장 널리 사용된다. 이 방법은 두 문서의 문장 비교에 있어 단어의 출현 순서를 고려하여 최장으로 일치하는 개수를 측정하여 두 문서의 유사성을 판단하는 척도로 같은 방법으로 작성된 두 문서의 비교에 응용할 수 있다. 그러나 본 논문에서는 제안하는 요약(추출 요약)과 평가에 이용하는 문서에 포함된 요약(생성 요약)이 서로 다른 기준으로 작성되어 ROUGE와 같은 방법을 그대로 응용 할 수 없다. 따라서 본 논문에서는 다음 식 (3)과 같이 두 요약에 출현한 단어들을 대상으로 ROUGE의 개발 근거인 precision(정확률)과 recall(재현율) 및 두 값의 조화평균인 F-measure를 함께 측정한다.

$$Precision = \frac{|Words_S \cap Words_R|}{|Words_R|}, \quad (3)$$

$$Recall = \frac{|Words_S \cap Words_R|}{|Words_S|}$$

$$F-measure = \frac{2 \times recall \times preision}{recall + preision}$$

여기서, $|Words_S|$는 시스템으로부터 생성된 요약에 출현한 중복되지 않은 단어의 수를, $|Words_R|$은 원 문서에 포함된 요약에 출현한 중복되지 않은 단어의 수를 그리고 $|Words_S \cap Words_R|$은 두 요약에 공통으로 포함된 중복되지 않은 단어의 수를 나타낸다. Precision은 시스템으로부터 생성된 요약이 불필요한 정보 없이 기존의 요약문과 얼마나 비슷하고 가깝게 생성되었는지에 대한 평가 값이며, recall은 생성된 요약이 기존의 요약문의 정보를 어느 정도나 포함하고 있는지에 대한 평가 값을 의미한다. 따라서 이 두 값이 모두 높게 계산되어야만 더 정확하고 적절한 양의 요약을 생성했다고 평가 할 수 있다. 그러나 precision과 recall은 계산의 특성상 서로 반대의 추세를 가진다. 즉, 시스템이 많은 양의 요약을 생성하는 경우 precision은 높게, recall은 낮게 계산되며, 시스템이 적은 양의 요약을 생성하는 경우 precision은 낮게, recall은 높게 계산된다. 따라서 이 두 값을 모두 고려하여 시스템을 평가하기 위해 두 값의 조화평균인 F-measure를 함께 측정한다[2].

4.2 실험 결과

제안한 요약 시스템의 객관적 성능 평가를 위해서는 기존의 요약에 가장 많이 사용되는 cosine 유사도에 기반한 문장 그룹화 방법과 함께 dice 유사도에 기반한 문장 그룹화 방법 및 top-n 문장 추출 방법과 함께 상용 프로그램인 MS-Word의 자동 요약 기능을 비교 평가 하였다. Dice 유사도는 비교할 벡터의 각 변수 값을 0 또는 1의 이진값으로 표현할 수 있을 때 효과적인 유사도 계산 방법이라고 알려져 있으며[7] top-n 문장 추출 방법은 문장 중요도 순으로 상위 n개의 문장만을 추출하는 가장 직관적인 추출 요약 방법이다. Cosine 및 dice 유사도의 그룹화를 위한 유사도 임계값은 실험을 통하여 0.7로 설정하고 그룹을 분할하여 그룹별로 중요 문장을 추출하였으며, 모든 방법에 대해 10%의 요약 비율을 적용하여 생성된 요약의 문장수가 원 문서 문장수의 10%를 넘지 않도록 하였다. 또한 모든 방법을 위한 핵심어 추출 및 핵심어의 중요도 계산은 [9]에서 사용한 공기 정보를 이용한 핵심어 추출 및 도합유사도를 이용한 핵심어 중요도 계산방법을 이용하였다. 본 논문에서 제안한 방법은, 그래프의 연결여부 결정을 위한 임계값에 따라 그 성능이 좌우될 수 있다. 따라서 연결 여부 임계값을 계산하는 기준이 되는 식 (1)의 n 값의 변화에 따른 다양한 실험을 통해 요약 성능을 평가하고, 적절한 n을 결정하여야 한다. 본 논문에서는 우선 제안한 방법의 타당성 검증을 목표로 하여 n이 1~3의 값을 가진다고 가정하고 요약 성능을 평가 하였으며 실험 결과 그림 3에 따라 n의 값이 2일 때의 결과를 기존의 요약 방법들과 비교 하였다.

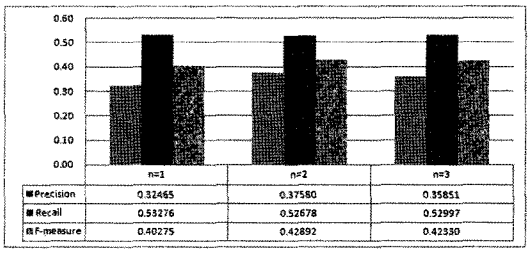
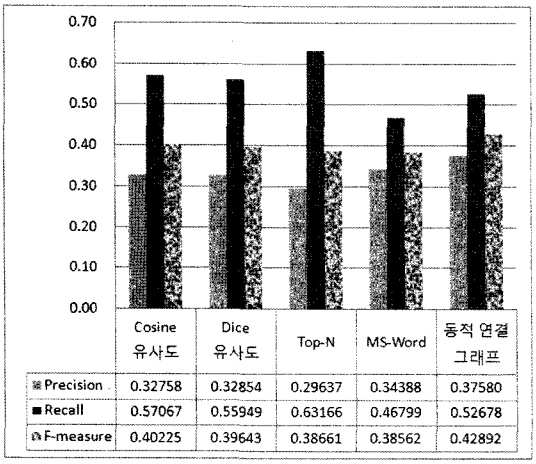


그림 3 그래프의 연결 여부 임계값의 계산 기준(n)에 따른 요약 성능 평가/비교



	Cosine 유사도	Dice 유사도	Top-N	MS-Word	동적 연결 그래프
Precision	0.32758	0.32854	0.29637	0.34388	0.37580
Recall	0.57067	0.55949	0.63166	0.46799	0.52678
F-measure	0.40225	0.39643	0.38661	0.38562	0.42892

그림 4 생성된 요약의 성능 비교/평가

기존의 주요 방법에 대한 요약 생성 성능 비교는 그림 4와 같다. 비교 결과, 제안한 시스템의 precision이 가장 높아 불필요한 정보 없이 기존의 요약문과 가장 비슷한 요약이 생성되었다고 평가할 수 있다. 그러나 recall은 다른 방법에 비해 상대적으로 낮은 값을 보이는데, 이는 다른 방법에 비해 비교적 적은 수의 문장을 요약으로 추출함으로써 기존의 요약을 상대적으로 적게 포함할 수밖에 없었기 때문인 것으로 판단된다. 종합적으로는 제안한 시스템의 F-measure값이 다른 시스템에 비해 높으므로, 제안한 요약 방법이 중요한 문장만을 잘 선별함으로써 전반적으로 요약에 좋은 성능을 보임을 알 수 있다.

5. 결론 및 향후 연구

본 논문에서는 실생활에서 발생할 수 있는 다양한 형태의 문서를 요약하기 위해 동적으로 문장 연결 그래프를 생성하고 이를 기반으로 하여 중요 문장을 추출하는 방법을 제안하였다. 특히 공통 포함 단어에 의한 문장간 연결 생성시, 문서내 문장들의 평균 길이에 따라 연결을 위한 최소 공통 포함 단어 수를 결정함으로써 실생활에서 발생할 수 있는 다양한 형태의 문서에 적합한 요약 기법을 구현하고, 기존 기술을 융합하여 응용 프로그램 문서로부터 요약을 자동으로 생성하는 시스템을 개발하였다. 또한 제안한 시스템의 객관적 평가를 위해 비교적 정확한 요약문이 실린 테스트 문서를 이용하여 성능을 평가하였다.

그러나 제안한 방법에서는 문장간 연결의 동적 결정을 위한 공통 포함 단어 수의 제한 값 결정에 직관적인 계산 방법을 이용함으로써 연결 여부 결정에 대한 객관적 근거가 미흡하다고 판단된다. 또한 본 논문에서 제안한 방법은 단지 다양한 문장 형태에 따른 문장간 연결 구조 파악만을 위한 방법으로, 문서 요약의 성능에 가장 크게 영향을 미치는 요인인 세부 주제별 유사 문장 그룹화 방법과 문장별 중요도 계산에 대한 개선은 고려하지 않았으며 성능평가에 사용한 문서 역시 공인된 데이타의 부재 문제로 인하여 학술지 논문만을 대상으로 함으로써 다양한 형태의 문서 요약에 대한 적합성 평가가 이루어지지 않았다.

향후에는 다양한 실험을 통하여 보다 객관적으로 문장의 연결 여부를 판단할 수 있는 근거를 마련함과 함께 보다 효율적인 요약문 생성을 위해 의미적 유사성에 기반하여 문장들을 그룹화 하고, 문장의 중요도를 계산하는 방법들에 대해 연구 할 것이다. 또한, 웹 문서나 뉴스 기사 등의 다양한 형태의 문서를 수집하고 이에 대한 공인된 요약을 생성, 수집하는 방법을 간구하여 실험 및 평가를 수행함으로써 제안한 방법이 학술지 논문과 같은 기술 문서에서 뿐만 아닌 실생활의 다양한 문서 요약에 적합한 방법인지를 검증할 것이다.

참 고 문 헌

[1] Inderjeet Mani, Automatic Summarization, Kohn Benjamins Publishing Co., 2001.

[2] Ohm Sornil, Kornnika Gree-ut, "An Automatic Text Summarization Approach using Context-Based and Graph-Based Characteristics," IEEE Conference on Cybernetics and Intelligent Systems, pp. 1-6, 2006.

[3] Daniel Mallett, James Elding, Mario A. Nascimento, "Information-Content Based Sentence Extraction for Text Summarization," IEEE International Conference on Information Technology: Coding and Computing, Vol.2, pp.214-218, 2004.

[4] Ani Nenkova, Lucy Vanderwende, Kathleen McKeown, "A Compositional Context Sensitive Multi-Document Summarizer: Exploring The Factors That Influence Summarization," Annual ACM Conference on Research and Development in Information Retrieval, pp.573-580, 2006.

[5] Takaharu Takeda, Atsuhiro Takasu, "UpdateNews: A News Clustering and Summarization System Using Efficient Text Processing," International Conference on Digital Libraries, pp.438-439, 2007.

[6] Il joo Lee, Minkoo Kim, "Document Summarization Based on Sentence Clustering Using Graph Division," Journal of Korea Information Processing Society, Vol.13-B, No.2, pp.149-154, 2006.

[7] Philipp Cimiano, Ontology Learning and Population from Text, Springer, 2006.

[8] Lei Yu, Jia Ma, Ren, F., Kuroiwa, S., "Automatic Text Summarization Based on Lexical Chains and Structural Features," IEEE ACIS International Con-

〔별지1〕 문서 요약의 실험에 사용된 문서명과 출처

일련번호	문서명(출처)
1	어절 내의 형태소 범주 패턴에 기반한 통계적 자동 띄어쓰기 시스템 (정보과학회논문지 : 소프트웨어 및 응용 제33권 제11호)
2	용어 발생 유사도와 퍼지 추론을 이용한 질의 용어 확장 및 가중치 재산정 (정보과학회논문지 : 소프트웨어 및 응용 제27권 제9호)
3	삭제된 파일 조각에서 기계어 코드 유사도를 이용한 악의적인 파일 탐지에 대한 연구 (한국정보보호학회 논문지 제16권 제6호)
4	인구 통계 정보를 이용한 협업 여과 추천의 유사도 개선 기법 (정보과학회논문지 : 컴퓨팅의 실제 제9권 제5호)
5	유사도 측정 기법을 이용한 효율적인 요구 분석 지원 시스템의 구현 (정보과학회논문지 : 소프트웨어 및 응용 제27권 제1호)
6	주요 색상의 분포 블록기호를 이용한 영상검색과 유사도 피드백을 통한 이미지 검색 (정보과학회논문지 : 소프트웨어 및 응용 제31권 제2호)
7	확장된 시퀀스 요소 기반의 유사도를 이용한 계층적 클러스터링 알고리즘 (한국 컴퓨터정보학회 논문지 제11권 제5호)
8	프로그램 유사도 평가 알고리즘 (한국 인터넷 정보학회 제6권 제1호)
9	유사도 분석과 명암 보정을 통한 혈관 추출 (한국 인터넷 정보학회 제7권 제4호)
10	SVM을 이용한 침입방지시스템 오경보 최소화 기법 (한국 인터넷 정보학회 제7권 제3호)
11	SVM 기반 히스테리시스 제어기를 이용한 D-STATCOM 전류 제어에 관한 연구 (전력전자학회 논문지 제11권 4호)
12	Single Nucleotide Polymorphism(SNP) 데이타와 Support Vector Machine(SVM)을 이용한 만성 간염 감수성 예측 (정보과학회논문지 : 시스템 및 이론 제34권 제7호)
13	센서 네트워크에서 효율적인 KNN 질의처리 방법 (정보과학회논문지 : 데이타베이스 제32권 제4호)
14	KNN 규칙과 새로운 특징 가중치 알고리즘을 결합한 패턴 인식 시스템 (전자공학회 논문지 제42권 CI 제4호)
15	SVM 기반의 효율적인 신분위장기법 탐지 (한국정보보호학회 제13권 제5호)
16	침입탐지를 위한 효율적인 퍼지분류규칙 생성 (정보과학회 : 소프트웨어 및 응용 제34권 제6호)
17	퍼지숫자를 기반으로 가중 구성요소를 갖는 퍼지시스템의 신뢰도분석 한국 인터넷 정보학회 제8권 제3호)
18	FCM 알고리즘과 퍼지 소속도를 이용한 지능형 자가 진단 시스템 (한국지능정보시스템학회 제13권 제1호)
19	앙상블 구성을 이용한 SVM 분류성능의 향상 (정보과학회논문지 : 소프트웨어 및 응용 제30권 제3호)
20	SVM 분류기를 통한 심실세동 검출 (전자공학회 논문지 제42권 SC편 제5호)

ference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Vol.2, pp.574-578, 2007.

[9] Il joo Lee, Minkoo Kim, "Multi-Document Summarization Based on Cluster using Term Co-occurrence," Journal of Korea Institute of Information Scientists and Engineers: Software and Application, Vol.33, No.2, pp.243-251, 2006.

[10] Chang-Beom Lee, Min-SOO Kim, Jang-Sun Baek, Hyuk-Ro Park, "Text Summarization using PCA and SVD," Journal of Korea Information Processing Society, Vol.10-B, No.7, pp.725-734, 2003.

[11] http://www.kings.co.kr, Kings Information & Networks.

[12] KLT 2.10b, http://nlp.kookmin.ac.kr/, Kookmin University.

[13] DBpia, http://dbpia.co.kr, 교보문고, 누리미디어

[14] Chin-Yew Lin, Franz Josef Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," Annual Meeting on Association for Computational Linguistics, No.605, 2004.

송 원 문

2004년 한신대학교 전자계산학과(학사) 2006년 숭실대학교 대학원 컴퓨터학과(석사). 2006년~현재 숭실대학교 대학원 컴퓨터학과 박사과정. 관심분야는 신경망, 음성인식, 데이타마이닝, 개인화

김 영 진

2004년 중부대학교 컴퓨터과학과(학사) 2006년 숭실대학교 대학원 컴퓨터학과(석사). 2006년~현재 숭실대학교 대학원 컴퓨터학과 박사과정. 관심분야는 기계학습, 데이타마이닝, 신경망, 에이전트 자율이동로봇

김 은 주

2001년 숭실대학교 자연과학대학 정보통계학과(학사). 2003년 숭실대학교 대학원 컴퓨터학과(석사). 2003년~현재 숭실대학교 대학원 컴퓨터학과 박사과정. 관심분야는 데이타마이닝, 기계학습, 웹마이닝, 추천 시스템, 신경망, 시맨틱 웹

김 명 원

1972년 서울대학교 응용수학과(학사). 1981년 University of Massachusetts(Amherst) Computer Science(석사). 1986년 University of Texas(Austin) Computer Science(박사). 1975년~1978년 한국과학기술 연구소 연구원. 1982년~1985년 Institute for Computing Science & Computer Application(Univ. of Texas). 1975년~1987년 AT&T Bell Labs. Member of Technical Staff. 1987년~1994년 한국전자통신연구소 책임 연구원. 1991년~1993년 충남대학교 전자계산학과 겸임부교수. 2000년~2001년 미국 IBM T.J WATSON 연구소 방문 과학자. 1994년~현재 숭실대학교 컴퓨터 학부 교수. 2002년~2003년 숭실대학교 정보지원처장. 2004년~2006년 숭실대학교 정보과학대학원장. 1992년~1993년 한국신경회로망 연구회 회장. 1992년~1993년 한국정보과학회 뉴로컴퓨팅 연구회 회장. 1993년~1995년 한국정보과학회 뉴로컴퓨팅 연구회 위원장. 1993년~1995년 IEEE Neural Network Council 한국지부장. 1998년~2000년 한국인지학회 부회장. 1997년~2000년 한국뇌학회 부회장. 2001년~2002년 한국뇌학회 회장. 관심분야는 유연주론, 신경회로망, 퍼지시스템, 진화알고리즘, 패턴인식, 자동추론, 기계학습, 데이타마이닝, creativity engineering 등