# 동적 연결 그래프를 이용한 자동 문서 요약 시스템[1)]

송원문○, 김영진, 김은주, 김명원
숭실대학교 컴퓨터 학부
{gtangel, liebulia, blue7786, mkim}@ssu.ac.kr

# Document Summarization System Using Dynamic Connection Graph

Won Moon Song○, Young Jin Kim, Eun Ju Kim, Myung Won Kim
Dept. of Computing, Soongsil University

## 1. 서론 및 연구배경

문서 요약이란, 보다 적은용량의 문서로 원본 문서에서 말하고자 하는 내용을 전달할 수 있도록 원본 문서로부터 가장 의미 있는 내용만을 파악하여 요약으로 구성하는 것을 말하며[1] 그 내용 구성 방법에 따라 생성요약과 추출요약으로 구분할 수 있다[2]. 특히, 추출요약은 원 문서로부터 중요하다고 생각되는 문장만을 선별하여 요약문으로 제공하는 것으로 생성요약에 비해 상대적으로 접근과 구현이 쉬워 많은 연구가 진행되고 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.

## 2. 동적 연결 그래프를 이용한 문장 연결 판단

문장간 연결 구조를 결정하는데 있어, [6]에서와 같이 단지 공통단어를 포함하고 있는지의 여부만 확인하여 연결 그래프를 구성하면, 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. 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 수식과 같이 문장의 평균 포함 단어 수에 따라 제한 값을 결정하고 문장의 연결 여부를 판단한다.

$$\theta_{Connection} = \frac{Average(|Word_{Sentence}|)}{n}, \quad CoOccur_{A,B} = |Words_{Sentence_A} \cap Words_{Sentence_B}|$$

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

$Average(|Words_{Sentence}|)$는 문서 내에 포함된 모든 문장의 평균 단어 수를 의미하며, $n$은 문장의 길이에 따라 $\theta_{Connection}$(공통 포함 단어 수의 제한 값)을 적절한 비율로 설정하기 위한 값으로 본 연구에서는 실험을 통하여 2로 정하였다. $|Words_{Sentence_A} \cap Words_{Sentence_B}|$는 임의의 두 문장 $A$와 $B$사이에 공통으로 포함된 중복되지 않은 단어의 수를 의미한다. 따라서, 이 값이 앞서 결정된 공통 포함 단어 수의 제한 값($\theta_{Connection}$) 이상인 경우에만 문장 $A$와 $B$는 그래프에서 연결된 것으로 표현한다.

## 3. 자동 문서 요약 시스템

실 생활에 문서 요약을 효과적으로 응용하기 위해서는 다양한 응용 프로그램 문서로부터 요약을 생성하여 사용자에게 제공하는 전과정이 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 1]과 같이 제안한 방법과 기존에 개발된 텍스트 추출 및 형태소 분석 프로그램과 그래프 분할 기법을 융합하여 다양한 응용 프로그램 문

1) 본 연구는 서울시 산학연 협력사업(10581cooperateOrg93111) | 결과로 수행되었음


그림 1. 자동 문서요약을 위한 프로세스

서로부터 자동으로 요약을 생성하는 자동화된 문서 요약 시스템을 구현하였다.

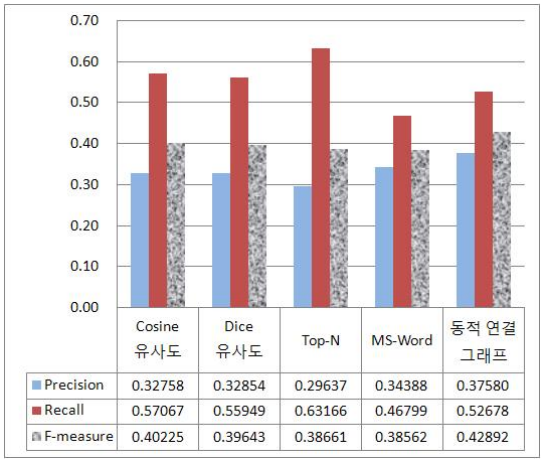
자동화된 문서 요약을 위해서는 가장 먼저 다양한 응용 프로그램 파일로부터 텍스트 정보만을 추출하며 이를 위해 관련 기술 보유 업체[7]의 텍스트 추출 제공 라이브러리를 제공받아 이용하였다. 이렇게 추출된 텍스트 정보로부터 형태소 분석기를 통해 명사나 동사, 형용사 같은 의미 있는 품사의 단어만을 따로 추출한다. 형태소 분석에는 연구용으로 공개되어 있는 국민대의 KLT[8]를 이용하였다. 형태소 분석을 통해 키워드의 후보가 되는 의미 있는 의미 있는 품사의 단어들이 추출되면 추출된 단어들의 공기 정보 분석[6]을 통하여 문서의 주요 키워드 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 그룹별 문장에 대해 포함된 단어들의 가중치를 고려하여 중요 문장을 추출한다. 추출된 문장을 문서내의 순서에 따라 정렬하면 사용자에게 제공할 요약 정보 생성이 완료 된다.

## 4. 실험 및 결론

제안한 요약 시스템의 성능 평가를 위해서는 국내의 전문 학술단체에서 발행한 IT분야의 학술지 논문 20건을 수집하여 이용하였다. 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. 문서 요약의 평가방법으로는 원래의 요약문서와 생성된 요약문서에 포함된 단어들을 대상으로 precision(정확률)과 recall(재현율) 및 두 값의 조화평균인 F-measure를 함께 측정한다.

$$Precision = \frac{|Words_S \cap Words_R|}{|Words_R|},\quad Recall = \frac{|Words_S \cap Words_R|}{|Words_S|},\quad F-measure = \frac{recall \times preision}{recall + preision}$$

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. 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. Cosine 및 dice 유사도의 그룹화를 위한 유사도 임계값은 실험을 통하여 0.7로 설정 하고 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. 각 방법에 대한 요약 생성 성능 비교는 [그림 2]에서 볼 수 있는바와 같이 다른 방법에 비해 제안한 요약 방법이 중요한 문장만을 잘 선별함으로써 전반적으로 요약에 좋은 성능을 보임을 알 수 있다.




| | 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 2. 생성된 요약의 성능 비교/평가

## 5. 참고 문헌

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

[2] Ohm Sornil, Kornnika Gree-ut, "An Automatic Text Summarization Approach using Content- 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 Sente-nce 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.

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

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