Journal of Korean Institute of Intelligent Systems, Vol. 22, No. 5, October 2012, pp. 656-662
http://dx.doi.org/10.5391/JKIIS.2012.22.5.656

LSA를 이용한 문장 상호 추천과 문장 성향 분석을 통한 문서 요약

Document Summarization Using Mutual Recommendation with LSA and Sense Analysis

이동욱* · 백서현* · 박민지** · 박진희* · 정혜욱* · 이지형*†
Dong-Wook Lee, Seo-Hyeon Baek, Min-Ji Park, Jin-Hee Park, Hye-Wuk Jung, and Jee-Hyong Lee†

*성균관대학교 컴퓨터공학과
**경희대학교 컴퓨터공학과

요 약

본 논문에서는 그래프기반 문장랭킹 방식인 문장 상호 추천과 문장의 주관, 객관 성향을 이용하는 문장 성향 분석을 혼합한 새로운 요약문 추출 방법에 대해서 기술한다. 문장 상호 추천에서는 문장을 단어벡터로 변환한 후에 LSA를 이용하여 문장과 문장 사이의 유사도 점수를 계산하였다. 이렇게 얻어진 유사도와 각 단어의 희귀도(Rarity Score)를 기반으로 문장과 문장 사이의 연결 강도를 정의하여, 그래프 기반 문장 랭킹 방식을 적용 하였다. 한편, 문장성향 분석에서는 주관, 객관 성향을 결정하기 위해서 기존의 Golden Standard 단어 성향 분류를 기반으로 워드넷을 확장하여 데이터베이스를 구축하였다. 이를 통해 각 단어들의 성향을 판단하고 단어들의 평균 성향을 문장의 전체 성향에 반영하여, 주관적 성향을 띄는 문장들을 선택하였다. 최종적으로 문장 상호 추천 결과와 문장 성향 분석 결과를 혼합하여 주어진 문서로부터 요약문을 추출하였다. 요약문 추출 기능의 객관적인 성능 평가를 위하여 추출된 요약문 토대로 한 분류게임을 실시하였고, 그 결과를 MS-Word에 포함된 문서 요약 기능과 비교함으로써, 제안한 모델의 효과성을 확인하였다.

키워드 : 문서 요약, 문장 상호 추천, Latent Semantic Analysis, 문장 성향 분석

Abstract

In this paper, we describe a new summarizing method based on a graph-based and a sense-based analysis. In the graph-based analysis, we convert sentences in a document into word vectors and calculate the similarity between each sentence using LSA. We reflect this similarity of sentences and the rarity scores of words in sentences to define weights of edges in the graph. Meanwhile, in the sense-based analysis, in order to determine the sense of words, subjectivity or objectivity, we built a database which is extended from the golden standards using Wordnet. We calculate the subjectivity of sentences from the sense of words, and select more subjective sentences. Lastly, we combine the results of these two methods. We evaluate the performance of the proposed method using classification games, which are usually used to measure the performances of summarization methods. We compare our method with the MS-Word auto-summarization, and verify the effectiveness of ours.

Key Words : Document Summarization, Mutual Recommendation, Latent Semantic Analysis, Sense Analysis

1. 서 론

신문, 책, 논문 등을 컴퓨터나 스마트기기를 이용하여 디지털 문서의 형태로 읽는 일은 현대 사회인이 지식을 얻기 위한 필수적인 활동이 되어가고 있다. 인터넷과 같은 가상공간에서의 디지털 정보량은 매우 빠른 속도로 증가하고 있기 때문에, 방대한 양의 정보 속에서 사용자가 원하는 정보만을 제공할 수 있는 컴퓨터 공학적인 정보 검색 기술이 꾸준히 요구되고 있으며, 그러한 기술 중에는 문서를 요약하거나 중요한 내용만 추출하는 기술도 포함된다.

어떠한 문서를 읽든 사람들은 글의 주제를 먼저 파악하는 방식으로 글을 이해하며 하나의 문서에서 얼마나 빠른 시간 안에 주제문을 도출하느냐 하는 문제는 정보의 생산성과 직결되는 문제이기도 하다. 따라서 현재 많은 정보 산업 분야에서 주제문 검출 및 문서요약

접수일자: 2012년 4월 20일
심사(수정)일자: 2012년 9월 29일
게재확정일자 : 2012년 9월 30일
† 교신 저자
이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No.2012-008062).

에 관한 연구가 활발히 진행 되고 있으며 다양한 방법을 이용한 문서요약 서비스나 엔진이 개발되어 있다.

그러나 문서의 종류, 분야, 분량, 작가의 문체 등 여러 가지 요소에 의해 문서의 특성이 매우 다양해지기 때문에 컴퓨터 알고리즘에 의해 높은 정확도로 주제문을 검출하는 것은 그리 간단하지가 않다.

주제 문장을 추출하는 대표적인 방법으로는 서로 유사한 문장들을 그래프 구조로 연결하고 연결된 유사도를 합산하는 그래프 기반 랭킹(Graph-based ranking)[1]방법이 있으며, 이 밖에도 단어 빈도수와 문서빈도수를 고려한 가중치들의 합산으로 문장의 점수를 결정하고 랭킹하는 방법, 문장의 특성을 분석하여 통계적으로 주제문이 될 수 없는 문단이나 문장을 제거하는 경험적 방법 등 다양한 공학적인 방식이 존재한다.

하지만 단어 가중치에 따른 점수 합산 방식은 한 문서 내에서의 단어 빈도수만 합산하는 경우 원래 흔하게 쓰이는 단어들에 의해 에러가 발생할 수 있기 때문에, 문서 빈도수(Document Frequency)를 고려해야 한다는 단점이 있으며, 그래프 구조를 이용한 유사문장 연결 방식에서는 유사도를 측정하는 과정에서 문장과 문장 사이의 단어-단어 공기성(Co-ocurency)을 고려하지 않기 때문에 의미, 개념적으로 연결된 요소를 반영하지 못하였다는 단점이 있었다.

본 논문에서는 문장을 구성하는 단어들의 중요도를 등장빈도수와 희귀도에 의해 정의하는 방법과 선형대수학적 해법인 LSA를 통해 의미, 개념적 공기성을 문장간 그래프 연결 가중치에 반영하는 새로운 상호 추천 방식(Mutual Recommendation)을 제안하고, 여기에 문장의 주관성향 또는 객관성향을 파악하는 방식을 혼합한 새로운 주제문 추출 모델을 제안하고자 한다.

2절에서는 주제문장을 검출하기 위한 절차와 몇 가지 기존 검출방법들에 대해서 소개한 뒤, 제안하고자 하는 방법과 관련된 기술들에 대해서 설명한다. 3절과 4절에서는 본 연구에서 제안하는 새로운 주제문장 검출 방법을 설명하고 5절에서는 제안한 방법의 성능을 실험을 통해 평가해본다.

2. 관련연구

2.1 주제문장 검출

컴퓨터 알고리즘을 이용해 어떤 문서의 주제문장을 추출하려면, 대부분의 경우 다음과 같은 순서를 따른다.

① 문서를 문장 단위로 분해하여 자료구조에 저장한다.
② 분해된 각 문장들을 주제문으로써 타당한지에 관한 점수(중요도)를 부여해준다. 이때 한 가지 이상의 일련의 함수들이 적용될 수 있다.
③ 득점한 점수에 따라 문장들을 정렬하고 고득점을 받은 문장들이 문서의 주제문으로 선택된다.

위와 같은 순서로 주제문을 추출할 때, 2번 단계에서 어떠한 기준과 방법들로 문장들의 중요도 점수를 부여할 것인지가 매우 중요한 문제이며 가장 핵심적인 부분이라고 할 수 있다. 여기서 문장들의 중요도를 결정하는 방법에는 경험적 방법, TF-IDF 가중치에 의한 방법, 그래프 기반 랭킹 방법 등이 있다.

경험적 방법(Empirical Method)

통계적 결과에 의하면 주제문이 아닌 문장들의 공통적인 특징이 관찰된다[2]. 따라서 이러한 특징을 가진 문장의 중요도를 감점함으로서 잘못된 주제문 추출 가능성을 낮출 수 있다. 예를 들어, “this”와 같은 대명사가 포함된 문장은 통계적으로 주제문일 확률이 낮으므로 감점할 수 있다.

경험적 방법은 상용 주제문 추출 모델에서 흔하게 사용되는 유용한 방법이지만, 단독으로 사용할 경우에는 주제문 추출이 거의 불가능하기 때문에 다른 방법과 혼합하여 함께 사용되며, 이러한 경우, 감점비율을 정의하기가 까다롭다.

TF-IDF 가중치에 의한 방법

TF-IDF(Term Frequency - Inverse Document Frequency)는 정보 검색 분야에서 단어의 중요도를 판단할 때 많이 사용되는 가중치로써, 문서 내 해당 단어의 빈도수(TF)와 전체 문서집합에서 해당 단어를 포함한 문서의 수의 역수(IDF)를 곱한 값이다.

TF-IDF 가중치를 이용하면 문서에 등장한 모든 단어의 중요도를 구한 뒤, 각 문장이 포함하고 있는 단어들의 중요도를 반영함으로서 문장의 중요도를 결정할 수 있다.

여기서 IDF 값을 구하기 위해서는 비교대상이 되는 전체 문서집합이 존재하여야 하기 때문에, 독립된 문서를 대상으로는 이 방법을 사용할 수 없다. 본 연구에서는 IDF 값을 대체할 수 있는 RS(Rarity Score) 값을 새롭게 제안하였다.

그래프 기반 랭킹 방법

그래프 기반 랭킹(Graph-based Ranking) 방법은 각 문장을 정점으로 하는 그래프 구조를 이용하는 방법이다. 그래프가 구성되면, 각 정점에 연결된 연결선의 가중치를 합산하여 해당 문장의 점수를 산출하고 랭킹을 구한다. 이러한 방법을 사용하면 각 문장들간의 연관관계를 통해 전체 집합에서의 중요도를 결정하기 때문에, 문서를 대표하는 문장을 찾을 수 있다는 장점이 있다.

R. Mihalcea가 제안한 그래프 기반 랭킹 알고리즘(Graph-based Ranking Algorithms) [1]에서는 문장과 문장의 유사도를 연결가중치로 사용하여 각 문장들의 중요도를 결정하였다. 이 때, S_a와 문장 S_b간의 유사도는 다음 식(1)과 같이 정의하였다.

$$Similarity(S_a, S_b) = \frac{|C_{ab}|}{\log(|S_a|) + \log(|S_b|)} \qquad (1)$$

여기서 $|C_{ab}|$는 두 문장에 공통으로 포함된 단어 셋의 크기이며, $|S_a|$와 $|S_b|$는 각 문장의 단어 셋 크기이다. 각 문장의 단어 셋에는 조사와 접속사와 같은 덜 중요한 품사의 단어는 제외될 수 있다.

이러한 방법을 이용하여 일정한 값 이상의 유사도를 가지는 연결 관계만을 표현하면 그림 1과 같은 그래프로 표시할 수 있다. 그러나 이와 같이 단어의 직접적인 포함여부로 유사도를 측정하는 경우, 의미적으로 내제된 유사관계를 고려할 수 없다. 또한 유사도 이외의 문장의 중요도를 결정하는 요소는 랭킹에 반영하지 못하였다는 단점이 있다.

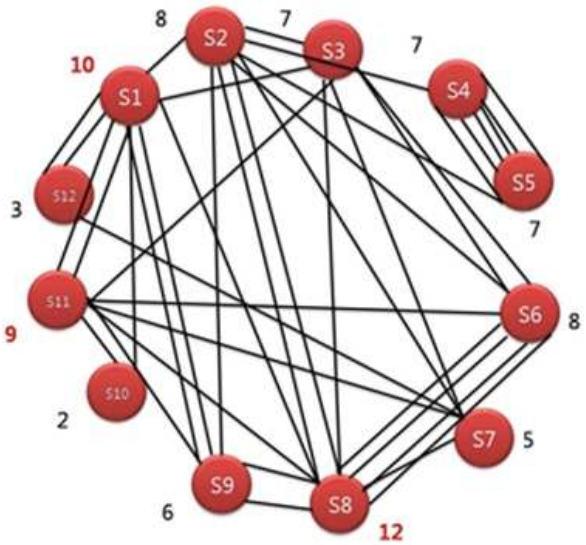

그림 1. 유사도 그래프 예시
Fig. 1. Example of similarity graph

본 연구에서는 위와 같은 그래프 기반 랭킹 방법을 사용하되, 연결 가중치 정의 방법을 변형하여 유사도의 측정방법을 개선하였고 유사도 이외의 요소들도 함께 고려할 수 있도록 하였다.

2.2 LSA와 워드넷(WordNet)

본 절에서는 제안하고자 하는 새로운 주제문 검출방법을 설명하기에 앞서서, 제안 방법에서 사용하는 LSA 기법과 유사 단어 확장을 위해 사용한 워드넷에 대하여 설명한다.

LSA(Latent Semantic Analysis)

LSA란, 단어와 문서의 분석에 쓰이는 선형 대수적 의미, 개념 분석 모델이다[3]. LSA의 input 으로는 각각의 문서들의 Stop word들을 제거한 word vector이며, 여러 문서의 word vector들의 Matrix를 SVD(Singular Vector Decomposition)을 이용하여, 단어와 문서의 축으로 이루어진 다차원 행렬을 3가지의 다른 행렬로 분해하게 되며, 단어와 문서가 다른 차원 축의 각각의 하나의 점으로 맵핑된다. 맵핑된 새로운 차원의 축들은 주어진 단어와 문서들의 정보를 잘 설명하는 축이 되며, 분산이 작은 축을 제거함으로써 차원을 축소하여 정보계산량 및 분석 속도를 줄일 수 있다.

특히, 새로운 차원상의 점들은 고유벡터를 구하는 과정에서 각 요소사이에 부정 방정식적 관계를 가지기 때문에, 단어-단어, 단어-문서, 문서-문서 간에 공기성(Co-occurency) 정보를 포함하고 있다. 이 정보는 LSA 상에서는 언어의 개념적인 유사 정보로써, 각각의 점들로 이루어진 단어벡터나 문서벡터들의 코사인 유사도를 계산하여 구할 수 있으며, 이와 같은 성질은 연관단어 검출 등의 기능에 사용된다.

본 연구에서는 본래의 LSA에서 문서단위로 사용되던 단어벡터를 단어들로 이루어진 문장들의 벡터로 구성하고, LSA를 사용하여 각 문장들 간의 코사인유사도를 구하였다. 이는 단어의 매칭만으로만 유사도를 측정하던 방법과 비교했을 때 더 의미적인 유사도를 구할 수 있다는 장점이 있다.

워드넷(WordNet)

워드넷[4]은 단어들을 쓰임새에 따라 'Synset'이라고 하는 유의어 집단으로 분류해 놓은 데이터베이스이다. 워드넷에는 Synset들이 상위, 하위, 등위 등의 관계로 서로 연결되어 있기 때문에 자연언어처리 분야 전반에서 유용하게 사용된다.

본 연구에서는 5장에서 소개할 문장 성향 분석 방법에서 성향이 확정된 단어세트인 Golden Standard 단어로부터 임의의 단어의 성향을 판별하기 위한 데이터베이스를 확장 구축하는데 사용되었다.

3. 문장 상호 추천

본 절에서는 앞서 소개한 그래프 기반 랭킹 방식(Graph-based Ranking)의 새로운 방법으로, LSA-유사도를 사용해 문장-문장 간 유사도를 정의하고, 단어 희귀도를 그래프 가중치에 반영하여 검출 정확성을 높일 수 있는 문장 상호 추천(Mutual Recommendation) 방법에 대해서 기술한다.

3.1 TF-RS (Term Frequency-Rarity Score)

TF-RS란 전체 문서 세트가 없는 경우에 기존의 정보 검색 연구 분야에 범용적으로 사용되는 TF-IDF(Term Frequency - Inversed Document Frequency)를 대체하여 단어의 중요도를 결정할 수 있는 값을 고안한 것으로써 본 연구에서는 IDF 값 대신 RS(Rarity Score, 희귀도) 값을 사용한다.

RS는 각종 문서에서 미리 수집된 10억 개의 단어 중에서 해당 단어의 등장한 횟수를 근거로 하여 각 단어의 희귀도를 측정한 값이다. 수집된 단어 중 최저로 측정된 단어의 빈도수를 F_m이라고 하고 측정하려는 단어 w의 빈도수를 $f(w)$라고 할 때 $RS(w)$ 값은 식(2)과 같이 정의한다.

$$RS(w) = \frac{F_m}{f(w)} \tag{2}$$

이 때 측정하려는 단어가 수집된 단어 집합에 존재하지 않는다면 RS값은 1로 정의한다.

수집된 단어 수가 많을수록 더 사실적인 빈도수를 계산할 수 있으며 Gutenberg Project에서 보유하고 있는 자료[5]에 의하면 빈도순위가 약 10,000위 정도에 이르면 그림 2와 같이 빈도수가 감소하는 폭이 매우 낮아

짐을 보였다.

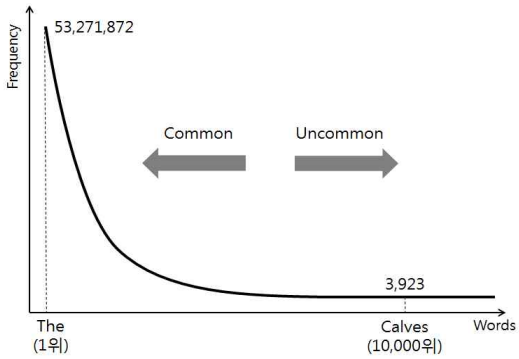

그림 2. 영어 단어 빈도수 [3]
Fig. 2. English words frequency per billion words [3]

본 연구에서는 문서 내에서 통계적으로 보았을 때 비교적 많이 등장한 단어가 무엇인지 TF-RS 값을 통하여 알아내고, 단어의 중요도를 결정하였다. 각 단어의 중요도는 3.3절에서 소개할 문장 상호추천에 반영하였다.

3.2 문장 상호 추천 (Mutual Recommendation)

그래프 기반 랭킹 방법에서 문장 간의 연결가중치를 LSA 코사인유사도로 정의할 수 있다. 이 방식의 장점은 직접적인 유사관계 뿐만 아니라 공기성(Co-occurency)에 의한 의미적 연결 관계까지 반영하여 개념적인 단계에서 문장과 문장의 유사도를 비교할 수 있다는 점이다.

그러나 LSA-유사도 그래프만 이용한 경우, 중요한 단어를 포함하지 않더라도, 문장 구조가 유사하거나 혹은 공통으로 포함하고 있는 단어가 많을 경우, 문장 상호 관련성이 높아질 수 있다.

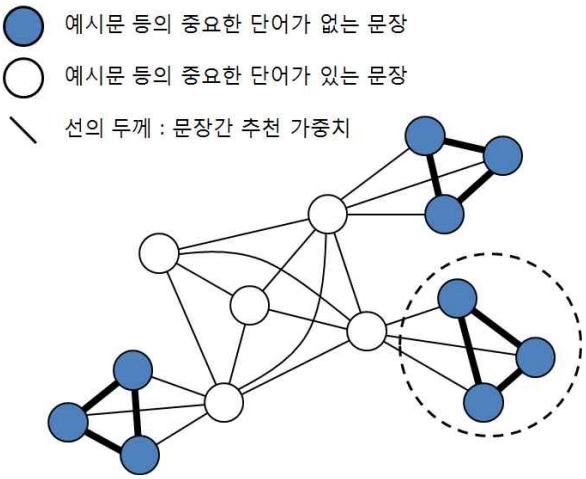

그림 3. 문장 그래프 상의 군집화 문제
Fig. 3. Problem in graph fragmentation

예를 들어 그림 3과 같이, 중요한 문장 5개와 중요도가 낮은 다수의 문장이 있다고 가정하자. 그림 3에서 각 원은 문장을 의미한다. 흰색 원은 주제문 등의 중요한 단어가 들어있는 문장을 뜻하고, 회색 원은 중요한 단어가 없는 문장을 뜻한다. 그래프 간선의 두께가 문장과 문장 사이의 관련성을 나타낸다.

만약 중요하지 않은 문장끼리 유사한 단어를 공통적으로 많이 포함하게 되면, 점선으로 표시된 부분과 같이 중요하지 않은 문장들 서로 간의 관련성이 높아지는 현상이 발생하게 된다. 이 경우, 전체적인 관점에서 해당 문장의 중요도가 높아지게 되어 최종적으로 중요한 문장보다 우선적으로 선택될 수 있다.

본 연구에서는 이 단점을 해결하기 위해서 문장-문장 간 유사도 점수에 연결된 각 문장을 구성하는 단어들의 TF*RS 기하평균을 곱하여 문장-문장 간 연결 가중치를 식 (3)와 같이 정의하였다.

$$W_{ij} = csim_{ij} \times \sqrt{TR(i) \times TR(j)} \qquad (3)$$

$$where\ TR(i) = \sum_{w \in i} (TF(w) \times RS(w))$$

여기서 $csim_{ij}$은 LSA에 의해 계산된 두 문장 i, j의 코사인 유사도를 뜻하며, $TR(i)$는 문장 i에 포함된 모든 단어들의 TF-RS 합산값을 뜻한다. 결과값인 W_{ij}는 문장-문장 간 상호 추천량을 뜻하며, 두 문장의 코사인 유사도에 TF-RS 합산값의 기하평균을 곱한 값이다. 이렇게 함으로서 중요한 단어가 포함된 문장들의 연결 가중치와 중요하지 않은 단어가 포함된 문장들의 연결 가중치의 편차를 높였고, 그 결과 군집화 문제를 해결할 수 있었다.

이렇게 만들어진 문장 상호 추천 그래프에서 각 문장의 중요도는 자신에게 연결된 가중치들의 합으로 결정된다. 전체 문장들을 높은 중요도 순으로 정렬하였을 때, 상위에 있는 문장들만 선택하여 요약문을 추출할 수 있다.

4. 문장 성향 분석

본 절에서는 문장 내 개별 단어들의 주관·객관 성향 수치의 평균을 문장의 주관·객관 성향 수치로 보고, 그 수치를 이용하여 문장의 중요도를 결정하는 방법을 설명한다. 이는 3장에서 소개된 문장 상호 추천과는 별개로 적용할 수 있으며 본 연구에서는 두 가지 방법을 적절한 비중으로 혼합할 것이다.

본 연구에서는 "문서의 주제문은 작가의 주관적인 생각이다"라는 사전적 정의를 토대로 주관적 성향이 강한 문장일수록 주제문일 확률이 높을 것이라는 가정을 하였다. 이에 따라, 주관적인 단어들이 많이 포함된 문장일수록 높은 점수를 가지도록 하였다.

단어의 주관성을 판단하기 위해서 Katja Markert의 성향 분석 연구 중 언어학자들에 의해서 단어의 주관•객관 성향이 확정된 Golden Standard (성향 확정 표준) [6] 단어 세트를 워드넷 상에서 확장 수집하므로써 데이터베이스화 하였다. 워드넷 확장은 기준 단어로부터 재귀적으로 상위어, 동의어로 연결된 단어들을 수집 하

였다.

워드넷을 확장한 데이터베이스는 다음과 같은 두 가지 특징을 반영하여 구축하였다. 구축결과 본래의 Golden Standard (성향 확정 표준) 단어에서 확장된 단어 일수록 중립적인 단어가 되는 경향을 보였다. 그림 4는 확장 방법을 의미한다. 그림 왼쪽 중앙의 검은 원은 'Happy', 'Glad' 등의 Golden Standard Word 중 Subjective의 성향을 띄는 단어를 의미한다. 그리고, 이 원을 기준으로 뻗어나간 화살표는 워드넷 상의 상위어와 동의어로 연결된 단어들을 재귀적으로 탐색하는 것을 뜻한다. 실험적으로 3단계 이상의 확장에서는 거의 모든 단어가 'Concept' 등의 중립적인 단어들이었다. 그림 4의 가장 오른편에 위치한 원들이 이러한 단어들을 의미한다. 이에 따라 각 단어는 2단계까지만 확장하였다.

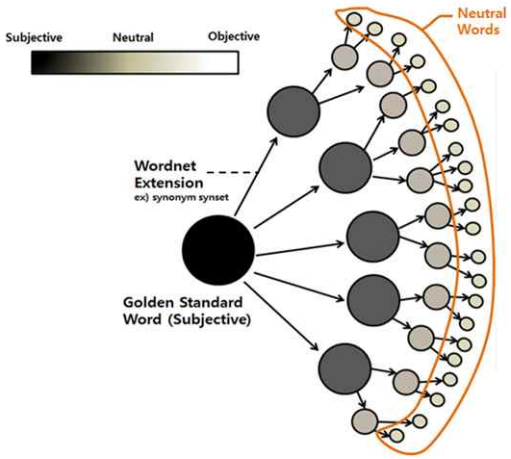

그림 4. 성향 확정 표준 주관 단어의 확장

Fig. 4. Extense of Subjective Golden Standard word

이렇게 확장된 단어의 주관·객관 성향 데이터베이스를 기반으로, 문장 내 단어에 주관·객관 점수인 SS(Sense Score)를 정의하였다. SS(Sense Score)는 주관적일수록 1에 가깝고, 객관적일수록 0에 가까우며, 중립인 경우 0.5가 되도록 정의하여 단어별 주관·객관도를 나타내도록 하였다. 다음 식(4)는 SS(Sense Score)의 값을 구하는 방법이다.

$$SS_n = SS_{n-1} \times \frac{1}{\sqrt{(2SS_0)}} \qquad (4)$$

SS_0는 기준 단어의 성향 점수이고, n은 탐색 단계를 의미한다. SS_0는 Golden Standard에 포함된 기준 단어의 주관·객관도를 의미하며, SS_n는 기준 단어에서 n차 확장된 단어의 주관·객관도를 의미한다. 식(4)의 의미는 워드넷 확장 시 재귀적인 각 단계를 탐색할 때 마다 성향점수가 하락 하거나 증가하여, 3단계 이상에서는 대부분의 단어가 중립적인 성향이므로 0.5점으로 수렴한다는 것을 의미한다. 워드넷 확장 시 중복으로 확장된 단어에 대해서는 SS(Sense Score)값의 평균을 내어 SS(Sense Score)의 값을 구하였다.

데이터베이스를 구축한 이후에는 각 문장의 개별 단어들을 데이터베이스에서 존재하는 여부를 판단한다. 각 단어 별로 존재하지 않는 단어에 대해서는 중립적인 단어로 판단하여 점수를 주고, 존재하는 단어에 대해서는 데이터베이스에 저장된 SS(Sense Score) 값으로 점수를 주었다. 문장의 성향은 문장 내 개별 단어들의 Sense Score 값의 평균을 내어 산출하였다.

본 연구에서는 문장 상호 추천 방식과 문장 성향 분석 방법을 적절한 가중치 비율로 계산하여 각 문장의 점수를 산출 및 랭킹하여, 상위 랭킹의 문장을 중요문장으로 선택하였다.

5. 실험 환경 및 결과

5.1 실험환경

실험 방식으로는 분류게임을 시도하였다. 분류게임이란, 1998년 5월 미국정부 TIPSTER Text Program (Phase3)에서 처음 시행한 대규모 요약기능 평가 시스템[7]으로서, 내용을 바탕으로 미리 카테고리가 분류된 문서들을 컴퓨터가 요약하면, 그 요약문을 사람이 읽고 원문의 카테고리를 분류한다. 만약 요약문만을 읽은 사람이 원문의 카테고리를 정확히 분류하면 요약문의 정확도가 높다고 할 수 있으므로, 분류게임에서 제대로 분류된 문서의 수가 많을수록 좋은 성능의 방법으로 판단한다.

본 연구에서는 TOFEL에서 52개, 수능에서 18개, 신문기사에서 31개, 총 101개의 문서를 추출하여 실험에 사용하였고, 한 문서는 정치, 경제 등 총 15가지의 카테고리 중 두 개의 카테고리에 포함되도록 분류하였다.

요약문은 각 경우의 문장순위 1위부터 3위까지 세 개의 문장을 단순 연결하여 생성하였고, 원문이 20문장 이하로 짧은 경우에는 전체 문장 수의 15%이하로 요약문을 추출하였다.

분류게임으로부터 문장 상호 추천 모듈과 문장 성향 분석 결과의 적절한 배합을 찾기 위하여 (문장 상호 추천 결과 : 문장 성향 분석 결과)의 반영 가중치를 (1.0:0.0), (0.3:0.7), (0.5:0.5), (0.7:0.3), (0.0:1.0) 과 같이 변화시켜가며 실험하였다. 예를 들어, (0.3:0.7)의 경우, 문장 상호 추천 모듈로부터 만들어진 문장 순위와 문장 성향 분석 모듈로부터 만들어진 문장 순위를 각각 백분율로 환산한 뒤, 3대 7의 가중치로 합한 값으로 최종 순위를 결정하였다.

각 가중에 대해 101개의 문서를 각 방법으로 요약문을 추출한 후에 5인의 전문가에게 분류게임을 의뢰하였다.

5.2 실험결과

분류게임에서 각 문서의 정답 카테고리 두 개를 모두 맞히는 것이 둘 중 하나를 맞히는 것보다 어렵기 때문에, 이를 반영하기 위하여 아래와 같은 수식을 기반으로 점수를 주었다.

$$Score(d, M) = 3^{C(M,d)} \qquad (5)$$

식 (5)에서 d는 문서이고, M은 요약 방법론이고, $C(M,d)$는 M이 추출한 요약문을 읽고 전문가가 맞힌

*d*의 카테고리 개수이다. 즉, 정답 카테고리 2가지를 모두 맞혔을 경우 9점, 카테고리 두 가지 중 1가지만 맞혔을 경우 3점, 모두 틀렸을 경우 1점이 점수로 주어지게 하였다.

제안하는 방법의 성능 비교를 위하여 MS Word 2007의 요약 결과와 비교하였다.

표 1. 실험 결과
Table 1. Result

	Proposed Method					MS Word
	1.0:0.0	0.7:0.3	0.5:0.5	0.3:0.7	0.0:1.0	
Point(%)	59.15	70.21	61.42	58.49	37.54	53.73
SetCon	5.05	5.71	5.79	4.85	2.96	6.02
DocCon	7.03	4.69	6.25	7.68	5.78	7.12

표 1은 문장 상호 추천 결과와 문장 성향 분석 결과를 반영할 가중치를 (1.0:0.0), (0.3:0.7), (0.5:0.5), (0.7:0.3), (0.0:1.0) 과 같이 변화시켜가며 실험한 결과이다. 여기서 Point(%)는 9점 만점 대한 상대적인 점수로, 모든 실험 횟수와 지문들의 평균 점수의 백분율 수치이다. 예를 들어 60%는 평균 5.4(=9*60%)점을 받았음을 의미한다. 즉, 높은 백분율 일수록 정확한 요약을 하였음을 의미한다. SetCon은 각 지문들의 점수 분산의 평균으로 실험 세트들의 일관성(Set Consistency)을 나타내며, 작은 값일수록 실험이 일관되게 진행되었다는 것을 의미한다. DocCon(Document Consistency)은 각 문서별 점수 평균의 분산으로 요약 품질의 일관성에 대한 실험결과를 나타내며, 이는 작은 값일수록 문서의 종류나 카테고리, 형태 등에 성능의 영향을 덜 받는다는 것을 말한다.

실험 결과, 표 1과 같이 문장 상호 추천의 비율이 문장 성향 분석 비율보다 높은 경우의 성능이 대체로 좋았으며, 혼합비율 (0.7:0.3)의 성능이 혼합비율 (1.0:0.0)보다 높은 것으로 볼 때, 문장 성향 분석을 포함 안하는 것 보다는 어느 정도 포함하는 것이 성능향상에 도움이 됨을 알 수 있다. 이는 글을 파악할 때, 단어들의 성향을 고려하는 것이 타당하다는 것을 의미한다. 하지만, 문장 성향 분석만을 요약에 이용한 혼합 비율 (0.0:1.0)의 실험결과에서, 가장 낮은 수준의 성능을 보인 것으로 보아 문장 성향 분석 방법만으로는 좋은 성능을 기대하기 힘들다는 것을 알 수 있다.

MS Word와의 비교에서도 (0.0:1.0) 경우를 제외한 네 가지의 경우에서 더 좋은 정확도와 일관된 성능을 보였다.

6. 결 론

본 논문에서는 그래프 기반 분석과 문장성향 기반 분석 방법론의 혼합한 새로운 문서 요약 방법을 제안하였다. 제안한 방법론의 분류 게임을 통하여 검증한 결과 기존상용프로그램인 MS Word 2007의 자동 문서 요약 기능 보다 나은 성능을 보임을 확인할 수 있었다.

분류 게임의 실험 결과를 보았을 때, 그래프 기반 분석 뿐 만 아니라 의미적 성향 분석을 혼합하여 분석하였을 때, 요약 성능이 향상 되었다. 이는 글을 파악할 때, 문장 상호 추천과 문장 성향 분석을 혼합한 방식이 문서 주제문 추출 및 요약에 유용함을 보여준다.

References

[1] R. Mihalcea, "Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization," *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, 2004.

[2] Jade Goldsteiny, *Summarizing Text Documents: Sentence Selection and Evaluation Metrics*, Language Technologies Institute Carnegie Mellon University, 1999.

[3] Scott Deerwester, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science*, 1990.

[4] G. A. Miller, "WordNet: An online lexical database," *Int. J. Lexicograph*, 1990.

[5] Word frequency list based on Project Gutenberg available at : http://en.wiktionary.org/wiki/Wiktionary:Frequency_lists, 2012.

[6] F. Su, K. Markert, *From Words to Senses: A case Study of Subjectivity Recognition*, School of Computing University of Leeds, 2008.

[7] 김영택 외, *자연언어처리*, 생능출판사, 2003.

저 자 소 개

이동욱(Dong-Wook Lee)
2006년~현재 : 성균관대학교 컴퓨터공학과 학사과정
관심분야 : 인공지능, 자연어처리
E-mail : replon@skku.edu

백서현(Seo-Hyeon Baek)
2006년~현재 : 성균관대학교 컴퓨터공학과 학사과정
관심분야 : 패턴인식, 기계학습
E-mail : blackskin2@skku.edu

박민지(Min-Ji Park)
2009년～현재 : 경희대학교
컴퓨터공학과 학사과정
관심분야 : 자연어처리
E-mail : pmjsky3265@gmail.com

박진희(Jin-Hee Park)
2011년～현재 : 성균관대학교
컴퓨터공학과 석사과정
관심분야 : 텍스트 마이닝, 정보검색
E-mail : gdiswarm@skku.edu

정혜욱(Hye-Wuk Jung)
1999년 : 한성대학교 정보전산학부
정보공학과 학사
2005년 : 성균관대학교 정보통신대학원
정보보호학과 석사
2007년～현재 : 성균관대학교 컴퓨터공학
박사수료

관심분야 : 패턴인식, 지능시스템, 기계학습
E-mail : wukj@skku.edu

이지형(Jee-Hyong Lee)
1993년 : 한국과학기술원 전산학과 학사
1995년 : 한국과학기술원 전산학과 석사
1999년 : 한국과학기술원 전산학과 박사
2002년～현재 : 성균관대학교 정보통신대
학 부교수

관심분야 : 퍼지이론, 지능시스템, 기계학습
Phone : 010-9265-0323
E-mail : john@skku.edu