The Journal of The Institute of Internet, Broadcasting and Communication (IIBC)
Vol. 16, No. 1, pp.33-42, Feb. 29, 2016. pISSN 2289-0238, eISSN 2289-0246

http://dx.doi.org/10.7236/JIIBC.2016.16.1.33
JIIBC 2016-1-5

대용량 오피니언 문서에 대한 특성 기반 요약 기법

Feature-Based Summarization Method for a Large Opinion Documents Collection

장재영*

Jae-Young Chang*

요 약 최근 SNS나 포털을 중심으로 다양한 분야 대해 대중들의 의견이 표현될 수 있는 환경이 확대되고 있고, 이로 인해 오피니언 문서들은 빠르게 대량화 되고 있다. 이러한 환경에서 대용량의 오피니언 문서들의 내용을 파악하기 위해서는 자동 요약 기술의 적용이 필수적이다. 하지만 오피니언 문서 내에는 대상 객체가 갖는 특성들과 주관적 표현들이 내재되어 있어 일반적인 요약 기법으로는 효율적인 요약이 불가능하다. 본 논문에서는 대용량의 오피니언 문서를 대상으로 주요 문장들을 추출하여 요약하는 기법을 제안한다. 제안된 기법에서는 사전에 정의된 오피니언 문서의 특성들에 대해서, 특성들에 대한 오피니언이 표현된 대표적인 문장들이 추출되도록 설계되었다. 또한 실험을 통하여 제안된 방법의 유용성을 증명하였다.

Abstract Recently, an environment in which public opinions are expressed about various areas is expanded around SNSs or internet potals, thus, opinion documents get bigger rapidly. Under these circumstances, it is essential to utilize automatic summarization techniques for understanding whole contents of large opinion documents. However, it is hard to summarize efficiently those documents with traditional text summarization technologies since the documents include subject expressions as well as features of targets objects. Proposed method in this paper defines features of opinion documents, and designed to retrieve representative sentences expressing opinions of those features. In addition, through experiments, we prove the usefulness of proposed method.

Key Words : Opinion Mining, Opinion Documents, Movie Reviews, Automatic Summarization, SNS

Ⅰ. 서 론

Web 2.0시대의 도래로 인한 SNS의 빠른 확산은 대중들이 사회 전반에 대한 그들의 주관적 의견(subject opinion)들을 피력할 수 있는 다양한 장의 토대가 되고 있다. 점차 대용량화 되고 있는 이러한 주관적 의견에 대한 문서들에 대해 자동 분류(automatic classification)나 검색(retrieval), 요약(summarization)에 대한 요구가 점차 늘고 있으며, 이러한 요구사항들이 오피니언 마이닝(opinion mining)의 발전을 가져왔다. 오피니언 마이닝은 주관적 문서로부터 작성자의 감정(sentiment)을 추출하여 다양한 분석을 시도하는 기술로서, 2000년대 이후 많은 연구가 이루어지고 있다[1-4]. 오피니언 마이닝에서의 가장 핵심적인 요소는 오피니언 문서(opinion document)가 해당 객체(object)에 대해서 긍정(positive) 혹은 부정(negative)적인 감정을 갖고 있는지 판단하는 감성 분석

*정회원, 한성대학교 컴퓨터공학과
접수일자: 2015년 11월 25일, 수정완료: 2016년 1월 12일
게재확정일자: 2016년 2월 5일

Received: 25 November, 2015 / Revised: 12 January, 2016 / Accepted: 5 February, 2016
*Corresponding Author: jychang@hansung.ac.kr
Dept. of Computer Engineering, Hansung University, Korea

(sentiment analysis) 기술이며, 이 기술에 대한 연구가 오피니언 마이닝의 대부분을 차지하고 있다. 감성 분석은 문서전체에 대해 전반적인 극성(polarity)을 결정할 수도 있고, 특성(features)별로 세분화하여 특성별 극성을 판단할 수도 있다. 이외에도 문서로부터 오피니언의 특성들을 자동적으로 추출하거나[5] 오피니언 문서에 대한 검색, 요약 등에 관한 연구도 이루어지고 있다[6]. 최근에는 트위터(twitter)와 같은 SNS 환경에서의 오피니언 마이닝에 관한 연구도 활발히 진행되고 있다[7].

대량으로 축적되고 있는 오피니언 문서들을 한눈에 파악하기 위해서는 감성분석 기술도 필요하지만, 분류된 문서들 중에서 핵심적인 내용을 추출해주는 요약 기술도 매우 중요하다. 자동요약 기술은 대용량 문서나 문서 집합으로부터 핵심 문서요소(text unit, 문장이나 어구 등)를 자동으로 추출하거나 구성하는 기술로 현재 많은 연구가 이루어지고 있다[8, 9, 10, 11, 12, 13]. 자동 요약 기술의 핵심은 문서요소들에 대해서 중요도를 판별하는 것인데, 이는 문서요소들의 출현 빈도나 문서요소 간의 유사도(similarity) 등을 정량적으로 계산하여 해결하고 있다. 그러나 오피니언 문서에 대한 요약은 기존의 자동 문서 요약 기술을 그대로 적용하기에는 한계가 있다. 우선 오피니언 문서는 긍정과 부정의 의미가 뚜렷하다. 따라서 긍정을 표현하는 문서와 부정을 표현하는 문서를 별도로 취급해야한다. 나아가 긍정과 부정을 나타내는 정도에 따라 다단계 클래스로 나눠 다룰 필요도 있다. 두 번째는 긍정과 부정의 표현이 하나의 객체 전반에 대한 표현일 수도 있지만 대부분은 객체가 갖는 내부 특성들에 대한 감성표현이다. 예를 들어 상품평과 같은 오피니언 문서에서 카메라 객체의 경우 렌즈, 화질, 디자인, 가격 등이 특성이 될 수 있고, 이러한 각 특성들에 대해 긍정 혹은 부정의 표현들로 구성된다. 따라서 오피니언 문서의 자동요약은 이러한 특성 위주로 추출되도록 고안되어야 한다.

본 논문에서는 대용량의 오피니언 문서를 대상으로 주요 문장들을 추출하여 요약하는 기법을 제안한다. 제안된 기법은 단문(short document)으로 구성된 오피니언 문서를 대상으로 하였으며, 문서집합 전체를 대표하는 개별 문서들을 추출하는 것을 목표로 하였다. 이를 위해 문서의 대표성을 판단하는 수식을 개발하였으며, 문서들을 중요도에 따라 랭킹을 하였다. 문서의 대표성은 문서의 길이, 문장 내에 언급된 특성의 비율, 전체문서에서 특성들의 출현빈도 등으로 계산하였다. 또한 오피니언 마이닝에서는 특성들의 정의가 매우 중요한데 본 논문에서는 이를 위해 정의된 특성들의 유사성을 계산하고 유사 특성들을 토픽(topic)으로 묶기 위한 방법도 제시하였다. 특히 토픽들의 유사성은 WordNet[14]을 활용하였으며, 유사 특성들을 묶어 토픽을 정의하기 위해 클러스터링(clustering) 기법을 활용하였다.

본 논문에서 제안된 요약 기법에 대해 실험을 실시하였다. 실험은 네이버 영화평을 대상으로 하였다. 네이버 영화평은 한글 140자 이하의 단문으로 구성되었으며, 수집하기가 용이하다. 또한 모든 영화평마다 평점이 부과되어 별도의 감성 분석 없이 문서들의 극성을 추정할 수 있다. 실험을 통하여 본 논문에서 제시한 요약 기법으로 선별된 문서들이 전체 문서를 얼마나 대표할 수 있는가를 평가하였으며, 이를 통하여 제안된 방법의 유용성을 증명하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해서 논하고 3장에서는 특성들의 정의와 토픽 생성 기법에 대해 설명한다. 4장에서는 문서를 중요도에 따라 랭킹하여 요약 문서를 선별하는 기법에 대해 설명하고 5장에서는 실험결과를 제시한다. 마지막으로 6장에서는 결론을 맺는다.

II. 관련연구

일반문서에 대한 자동요약은 예전부터 많은 연구가 진행되었다. 자동요약 방식은 크게 문서 내에서 문장이나 구(phase)같은 중요한 문서요소(text unit)들을 추출(extraction)하는 방식이 있고, 문서요소를 새롭게 생성하여 요약하는 방식(abstraction)[15]이 있다. 그런데 후자의 방식은 너무 어려워 대부분의 자동요약 연구는 전자의 방식에 초점이 맞추어져 있다. 또한 요약대상에 따라 하나의 문서로부터 요약하는 단일문서 요약(single document summarization) 방식[16]과 여러 개의 문서로부터 요약하는 다중문서 요약(multi-document summarization)방식[9, 10]으로 나눌 수도 있다. 단일문서 요약은 하나의 문서 내에서 해당 문서를 대표하는 주요 문서요소를 추출하는 것으로 비교적 목표가 단순하지만, 다중문서 요약의 경우에는 대표되는 문서요소를 추출하는 것뿐만 아니라, 여러 개의 문서에 산재한 다양한 주제들을 요약 정보에 모두 포함해야하는 문제도 안고 있다.

하지만 다중 문서들이 모두 하나의 주제에 대한 내용을 담고 있다면 여러 개의 문서를 하나의 큰 문서로 취급해 단일문서 요약 방식을 그대로 적용할 수도 있다.

가장 대표적인 문장요약 기법은 그래프(graph) 기반의 textRank[8] 기법이다. 여기서는 문장을 하나의 노드(node)취급하고 노드 간의 간선(edge)은 문장 간의 유사도에 따라 가중치(weight)를 부여하여 구성한다. 이렇게 구성된 그래프에 대해서 PageRank[17]에와 유사한 방식으로 노드들을 랭킹하여 상위 N개의 노드(문장)들을 요약 결과로 최종 선택하게 된다. 이 방식과 같이 그래프를 이용한 요약은 비교적 단순하면서도 효율적인 것으로 알려져 있어, 많은 후속 연구[9, 10, 11]에서도 이 방식을 응용하고 있다. 또한 많은 연구에서 문서내의 내용을 여러 개의 소주제로 나누어 소주제별 대표 문장들을 선택하는 요약 방식을 제안하였다[12, 13]. 일례로 [12]에서는 연관단어들의 집합으로 클러스터를 구성한 후 각 문서와 클러스터 간의 유사도를 계산하여 클러스터를 대표하는 문장들을 랭킹하여 요약정보를 생성하는 방식을 사용하였다.

오피니언 문서에 대한 요약 기법은 비교적 연구가 많지 진행되지 않은 상태이다. 오피니언 문서의 요약은 일반 문서에서와는 다르게 특정 대상 객체에 대해 특성별로 긍정과 부정 표현에 대한 내용을 요약해야한다. 따라서 일부 연구에서는 (특성, 오피니언)의 쌍으로 구성된 형태로 요약 결과를 제공한다[18, 19, 20]. 이러한 결과를 제공하기 위해서는 문장으로부터 특성을 추출하고, 해당 특성에 대한 오피니언을 인식해야한다. 또한 인식된 오피니언에 대한 극성을 파악하고 그 결과를 통계나 테이블 형태로 제공하게 된다. 또 다른 형태는 특성과 오피니언을 기술한 문장을 추출하여 문장 단위로 요약결과를 제공하는 방식이다[21, 22]. 대표적으로 [21]에서는 트위터를 대상으로 토픽별로 오피니언이 포함된 게시글들을 수집하고 이중에서 대표적인 게시글을 대표문서로 선정하는 방식을 사용하였다. 대표 게시글 선택은 다른 게시글과 중복도, 토픽과의 연관정도, 문법적 완성도 등을 기준으로 삼았다.

Ⅲ. 특성 정의와 토픽 선정

오피니언 문서는 객체에 대한 전반적인 감성표현도 있지만 대부분은 객체의 세부 특성에 대한 감성표현이 대부분이다. 예를 들어 상품평에서 카메라 객체의 경우 렌즈, 화질, 디자인, 가격 등이 특성이 될 수 있고, 영화평의 경우에는 스토리, 감독, 배우, CG 등이 특성이 될 수 있다. 이러한 특성들은 객체마다 다르고, 동일 특성임에도 불구하고 다양한 표현방식이 사용될 수도 있다. 따라서 오피니언 마이닝에서 특성의 정의는 매우 중요한데, 많은 연구에서 오피니언 문서로부터 특성들을 자동으로 인식하여 정의하거나, 동일하거나 유사한 의미를 갖는 특성들을 클러스터링하여 소주제인 토픽으로 정의하는 연구가 이루어지고 있다[20, 23].

본 논문에서는 네이버 영화평과 같이 단문으로 구성된 오피니언 문서 집합으로부터 전체를 대표할 수 있는 문서를 선별하는 요약 방식을 제안한다. 본 논문에서는 전체적으로 영화평을 대상으로 한 요약 과정을 제시한다. 이를 위해서는 앞서 언급한 바와 같이 요약 대상 분야의 특성과 토픽을 정의해야한다. 영화평에 대한 특성은 비교적 분명하다. 예를 들어 [24]에서는 사전에 정의된 특성들을 그대로 사용하였다. 본 논문에서도 사전에 정의된 특성들을 활용하였다. 특히 영화의 경우에는 '감독', '스토리', 영상' 등과 같은 영화의 일반적인 특성뿐만 아니라 장르별 특성 그리고 출연자 및 배역 이름 등 개별 영화마다 정의할 수 있는 특징들도 존재한다. 따라서 이러한 특성들을 수작업으로 정의하였다. 예를 들어 2013년에 개봉한 영화 '은밀하게 위대하게'를 대상으로 선별한 특성들은 표 1과 같다.

표 1의 특성들을 보면 많은 특성들이 의미가 같거나 유사한 것을 알 수 있다. 따라서 유사한 특성들을 묶어 몇 개 단위의 토픽들로 클러스터링할 필요가 있다. 이렇게 되면 특성 위주의 요약이 아닌 토픽단위의 요약 결과를 생성할 수 있다. [24]에서는 토픽 생성에 있어서 수작업을 활용하였고, [20]에서는 특성과 감성단어와의 관계를 활용하여 클러스터링하는 방법을 활용하였다. 본 논문에서는 WordNet을 활용하여 사전적 유사성으로 클러스터링을 시도하였다. 다만 WordNet은 영단어에 대해 구축된 것이므로, 한글로 된 특성들을 영단어로 변역한 후 번역된 단어를 이용하여 WordNet에서 단어간의 유사도를 정량적으로 측정하였다. 이렇게 계산된 특성간의 유사도를 이용하여 클러스터링 알고리즘을 적용하고 최종적으로 표 2와 같은 토픽들을 선정하였다. 다만 표 2의 특성 중에서 배우 이름과 같은 고유명사는 자동 클러스터링이 불가능하므로 인위적으로 해당 클러스터에 배정

하였다.

표 1. 수작업으로 수집된 영화 '은밀하게 위대하게'의 특성 리스트
Table 1. Feature list of Movie 'Secretly Greatly' collected manually

웹툰 원작 내용 스토리 액션 전개 결말 여운 이야기 구성 엔딩 반전 코미디 완성도 시나리오 장르 줄거리 스토리라인 배경 감독 연출 작품성 작품 감정 편집 영상 인상 제작 장철수 감성 예술 노래 제작진 음향 스텝 음악 예술성 구성력 김수현 연기 배우 이현우 박기웅 캐 스팅 연기력 손현주 캐릭터 주인공 조연 원류환 인물 주연 고창석 리해랑 리해진 출연진 이채영 표정 박혜숙 김태원 출연자

표 2. 클러스터링을 이용한 토픽 생성 결과
Table 2. Topics Generated with Clustering

토픽	특성 리스트
스토리	웹툰 원작 내용 스토리 액션 전개 결말 여운 이야기 구성 엔딩 반전 코미디 완성도 시나리오 장르 줄거리 스토리라인 배경
연출	감독 연출 작품성 작품 감정 편집 영상 인상 제작 장철수 감성 예술 노래 제작진 음향 스텝 음악 예술성 구성력
배우	김수현 연기 배우 이현우 박기웅 캐스팅 연기력 손현주 캐릭터 주인공 조연 원류환 인물 주연 고창석 리해랑 리해진 출연진 이채영 표정 박혜숙 김태원 출연자

다음 단계는 수집된 각 영화평 문서들에 대해서 언급된 특성들을 추출하고, 해당 특성에 대한 오피니언 극성(긍정/부정)을 결정하는 것이다. 각 문서에서 언급된 특성을 추출하는 것은 단순 패턴매칭(pattern matching)을 이용하던지 형태소 분석기를 이용하면 된다. 다만 추출된 특성에 대한 극성은 감성 분류를 통해서만 판단할 수 있다. 하지만 본 논문에서는 별도의 감성분류 작업을 가정하지 않았다. 대신에 각 영화평에서 언급된 특성들에 대한 극성은 해당 영화평의 전체 극성과 같다는 휴리스틱(heuristics)을 사용하였다. 예를 들어, '연기'라는 특성이 평점 10점의 영화평에 언급되어 있다면, 그 영화평에서 '연기'라는 특성에 대해서는 긍정적 감성을 갖는다고 추정할 수 있다. 이와 같이 가정 하에 특성 혹은 토픽별로 각 극성에 대한 영화평들의 분포를 확인할 수 있다. 본 논문에서는 네이버 영화평에서 영화 '은밀하게 위대하게'를 대상으로 평점 1점과 평점 10점 영화평을 각각 2,940개 5,555개를 수집한 후에 각 영화평으로부터 표 2에서 정의된 특성들을 추출하였다. 그런 다음 각 특성에 대해서 빈도를 계산하였다. 표 3은 그 중에서 '스토리' 토픽에 속한 특성들의 분포를 보여준다. 이 표에서 평점 1은 부정 영화평, 평점 10은 긍정 영화평으로 간주하며, 앞서 언급한 대로 각 문서에서 언급된 특성도 동일한 극성을 갖는다고 가정한다. 이 표에서 보는바와 같이 '스토리' 토픽에서 가장 많이 언급된 특성은 '웹툰'이며 언급 횟수는 각각 1454번이다. 또한 평점 10과 평점 1의 영화평 중에서 '웹툰'이 언급된 영화평의 수는 각각 1061, 393으로 '웹툰'에 대한 긍정 영화평이 부정 영화평 보다 대략 3배 정도 많다는 것을 알 수 있다.

표 3에서 각 특성에 대한 문서 수는 요약을 위한 대표 문서를 선별하는 과정에서 중요하게 이용된다. 그 이유는 요약될 영화평들을 선정할 때 이 값들의 비율에 따라 결정해야하기 때문이다. 본 논문에서는 특성 f에 대한 이 값을 특성랭킹포인트(feature ranking point) $r(f)$로 정의한다. 또한 토픽 A에 포함된 모든 특성들에 대한 특성랭킹포인트의 합을 토픽랭킹포인트(topic ranking point) $r(A)$로 정의한다. 예를 들어 표 3에 대해서 특성 '웹툰'에 대해 $r(\text{웹툰})$은 1454이며, 토픽 '스토리'에 대한 $r(\text{스토리})$는 3850이 된다. 또한 긍정과 부정 영화평들을 분리하여 다루기 위해 특성 f에 대한 긍정과 부정 특성랭킹포인트를 각각 $r_{pos}(f)$, $r_{neg}(f)$로 정의한다. 예를 들어 $r_{pos}(\text{웹툰})$, $r_{neg}(\text{웹툰})$은 각각 1061, 393이다.

Ⅳ. 요약문서 추출 기법

본 논문의 목표는 표 3의 정보를 바탕으로 각 토픽을 대표하는 영화평들을 선별하는 것이다. 이를 위해서는 본 논문에서는 다음과 같은 기준으로 대표 영화평들을 선별하였다.

1. 주어진 토픽과 연관성을 강한 영화평
2. 길이가 긴 영화평
3. 토픽 내의 특성 중에서 출현빈도가 높은 특성에 대한 영화평

첫 번째 기준을 반영하기 위해서는 각 영화평 문서와

표 3. '스토리' 토픽의 각 특성에 대한 출현 빈도수
Table 3. Features Frequency in Topic 'Story'

특성	평점1(부정)		평점10(긍정)		전체	
	문서수	비율(%)	문서수	비율(%)	문서수	비율(%)
웹툰	393	33.45	1061	39.66	1454	38.63
원작	155	4.03	472	12.26	627	16.39
내용	166	4.31	317	8.23	483	12.66
스토리	144	3.74	197	5.12	341	8.95
액션	76	1.97	210	5.45	286	7.48
전개	68	1.77	78	2.03	146	3.84
결말	29	0.75	78	2.03	107	2.80
여운	12	0.31	88	2.29	100	2.61
이야기	23	0.60	33	0.86	56	1.47
구성	26	0.68	19	0.49	45	1.19
엔딩	8	0.21	38	0.99	46	1.20
반전	11	0.29	28	0.73	39	1.02
코미디	9	0.23	22	0.57	31	0.81
완성도	18	0.47	11	0.29	29	0.77
시나리오	18	0.47	4	0.10	22	0.58
장르	9	0.23	9	0.23	18	0.47
줄거리	10	0.26	8	0.21	18	0.47
스토리라인	0	0.00	1	0.03	1	0.03
배경	0	0.00	1	0.03	1	0.03
합	1175	100%	2675	100%	3850	100%

토픽간의 연관정도 정량적으로 계산하여 그 값으로 영화평들을 랭킹해야한다. 영화평 문서 s와 토픽 A간의 연관정도는 다음의 수식으로 계산할 수 있다.

$$d(A,s)=\frac{|F(A)\cap sf(s)|}{|sf(s)|} \quad (1)$$

여기서 $F(A)$는 토픽 A에 포함된 특성의 집합을 나타내며, $sf(s)$는 문서 s에서 언급된 특성들의 집합을 나타낸다. 따라서 $d(A,s)$가 1에 가까울수록 문서 s는 토픽 A에 대한 영화평임을 강하게 시사한다.

두 번째 기준인 영화평의 길이를 기준으로 삼은 이유는 단문으로 구성된 영화평의 특성을 반영한 것이다. 본 논문에서 실험 대상으로 선정한 네이버 영화평은 140자 이하의 단문만을 허용한다. 따라서 많은 영화평들이 매우 짧은 코멘트로 구성되어 있어 요약 대상으로는 적절하지 않은 것들이다. 따라서 이러한 영화평들을 배제하기 위해서 영화평의 길이를 선별 기준으로 채용하였다.

세 번째 기준을 사용한 이유는 동일한 토픽내의 특성들이라 할지라도 그 의미가 차이가 나는 경우가 많아, 되도록 자주 언급되는 특성위주로 요약 문서를 선정하는 것이 대표성을 지닐 수 있기 때문이다. 이 기준은 표 3의 특성랭킹포인트를 이용하면 곧바로 적용이 가능하다.

이와 같은 세 가지 기준에 적용하여 토픽 A를 대표하는 문서는 다음의 수식을 만족하는 문서 s가 된다.

$$argmax_{s\in S}(d(A,s)\times length(s)\times \sum_{f\in A\cap sf(s)} r(f)) \quad (2)$$

여기서 S는 전체 영화평 문서집합이고, $length(s)$는 문서 s의 크기를 나타낸다. 요약정보를 생성할 때 특정 토픽에 대해 식 (2)를 이용하여 문서들을 랭킹한 후, top N개를 최종 요약결과를 결정하면 된다. 그러나 N개의 요약 정보를 생성할 때 랭킹 결과를 그대로 생성할 경우에는 $r(f)$가 가장 높은 특성에 대한 영화평만 일방적으로 결정될 가능성이 높다. 따라서 하나의 요약 문서를 결정할 때마다 일방적으로 하나의 특성에 대한 영화평만이 나오지 않도록 $r(f)$와 $r(A)$를 매번 재조정해야한다. 이 방식을 이용한 요약문서 선택 알고리즘은 그림 1과 같다.

이 알고리즘은 주어진 토픽에 대해서 n개의 요약 문서를 선택하는 과정을 보여준다. 3번째 라인이 바로 요약문서를 선택하는 부분이며, 6번째 라인에 있는 내포된 순환문이 $r(f)$와 $r(A)$를 재조정하는 부분이다. 여기서 $r(f)$는 매번 $\frac{r(A)}{n}$만큼 그 비율만큼 재조정되어 동일한 특성에 대한 영화평이 반복적으로 선택되는 것을 방지하며, β값을 조정함으로 써 그 강도도 조정이 가능하

Algorithm SUMMARY_GENERATION(A, n)
Input A: topic
n: the number of documents to be selected
output summaryDoc: selected n documents
summaryDoc = {}
for(count=1; count<=n; count++) {
select $s \in S$ such that
$argmax_{s \in S}(d(A,s) \times length(s) \times \sum_{f \in A \cap sf(s)} r(f))$
summaryDoc = summaryDoc $\cup \{s\}$
for $\forall f \in F(A) \cap sf(s)$ {
$r(f) = r(f) - \left(\frac{r(A)}{n} \times \frac{\beta}{|A \cap sf(s)|}\right)$
$r(A) = \frac{n-1}{n} r(A)$
}
}
return summaryDoc

그림 1. 요약문서 선택 알고리즘
Fig. 1. Summary Document Selection Algorithm

다. $r(A)$ 또한 같은 방식으로 재조정된다. 최종적인 요약문서는 summaryDoc이라는 변수에 저장되어 반환된다. 그림 1에서 보여준 알고리즘은 긍정과 부정 영화평을 통합해서 요약하는 과정을 보여준 것이다. 긍정과 부정 영화평을 분리해서 처리하려면 전체 영화평을 긍정/부정 영화평 집합을 분할하고, 각각에 대해 특성랭킹포인트를 $r_{pos}(f)$와 $r_{neg}(f)$를 적용하면 된다.

알고리즘을 실행하여 얻은 요약문서의 일부는 그림 2와 같다. 이 그림은 긍정 영화평 중에서 '스토리' 토픽에 대해 실행한 요약 결과 중 상위 4개의 문서이다. 이 그림에서 보는 바와 같이 대부분의 요약결과가 표 3의 상위특징들에 대한 내용으로 구성된다는 것을 확인할 수 있으며, 가장 빈도수가 높은 '웹툰'에 대한 언급뿐만 아니라 그 다음 상위 빈도수를 차지하는 '원작', '내용' 등이 언급된 영화평도 선정되는 것을 확인할 수 있다. 이 결과는

1) 웹툰원작영화들이 스토리에 변화를 주었다가 실망한 경우가 많아서 걱정하였는데 매우 원작 스토리에 충실하여 만족스러웠습니다 캐스팅이 매우 적절하였는데 특히 손현주 님의 캐스팅이 신의 한수였다고 생각합니다
2) 웹툰보고나서보니 더 좋았던 부족한 스토리는 원작을 떠올리게해서 집중하게 만들고 오히려 몰입도 차라리 잘됨 그리고 스토리와 구성에서 미흡해 보이긴 하지만 그리 욕먹을 수준은 아니다 그리고 중간에 보고도 억지감동 나발부는 사람들은 영화를 제대로 안본거다
3) 캐스팅 잘한영화 김수현을 캐스팅한건 정말 잘한일이다 김수현이 캐스팅된게 아니였다면 망한영화들중 하나였을테지만 캐스팅이 김수현이되서 김수현 얼굴이라도 보러가는 영화원작웹툰을 뛰어넘지 못하고 후반부에서 웹툰을 미리 읽지않은 사람이라면 그저 고개만 갸웃할듯
4) 원작을 그대로 살리려고 노력은 하셨는데 디테일이랑 이야기흐름이랑 좀 안맞는달까 원작에서 내용을 잘라낸부분이있어서 좀 이쉽네요 근데 확실히 웹툰내용이 재밌어서 영화도 재미있네요 아직 원작안보신분들은 꼭 웹툰보세요

그림 2. 긍정영화평 최상위 요약결과
Fig. 2. Most Significant Summary Results of Positive Reviews

그림 1의 알고리즘에서 7번째 라인의 β값을 5로 설정한 결과이다. 앞서 설명한 바와 같이 β값에 따라 출현빈도수가 높은 특성을 언급한 문서가 일방적으로 요약 문서로 선택될지, 아니면 되도록 다양한 특성에 대한 영화평들이 골고루 선택될지가 결정된다. 이에 대한 실험 결과는 다음 장에서 다룬다.

V. 성능 평가

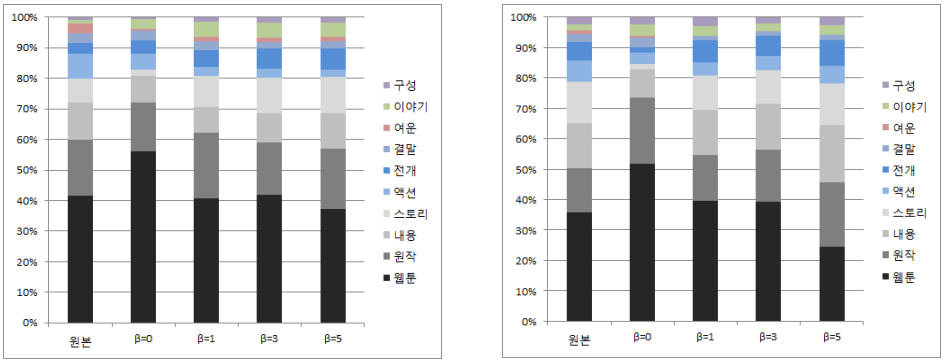
문서요약은 자동분류와 같은 다른 텍스트 마이닝 분야와는 다르게 실험평가의 객관성을 보장하는 것이 용이하지 않다. 그 이유는 테스트 문서에 대한 정확한 요약결과인 gold standard를 정의하는 것이 매우 어렵기 때문이다. 일반적인 문서요약 연구에서 많이 채용하는 실험대상은 논문이다. 논문은 앞부분에 초록이 있으므로 요약 결과가 초록과 얼마나 유사한가를 평가함으로써 요약결과의 정확성에 대한 평가가 가능하다. 그러나 오피니언 문서를 대상으로 한 요약은 그러한 비교대상이 존재하지 않는다. 이 경우에는 설문 조사를 통해 수작업으로 평가를 하는 수밖에 없는데 이 또한 객관성이 떨어지게 된다.

본 논문에서는 요약으로 선택된 문서들이 전체를 어느 정도 대표하느냐를 평가하기 위해 다른 접근방법을 채택하였다. 본 논문에서 대상으로 하는 문서들은 단문 오피니언 문서이며, 이러한 문서들은 대부분 대상 객체(영화)의 특성위주로 쓰여 있다. 따라서 요약으로 선택된 오피니언 문서들에서의 출현하는 특성 분포가 원본 문서들에서의 특성 분포와 유사하면 요약문서들은 원본문서들을 대표한다고 볼 수 있다. 이러한 관점에서 본 논문에서는 원본 문서들과 요약결과 문서들에서 특성들을 각각 추출하여 그들의 분포를 비교하였다. 이를 위해 앞선 예에서와 같이, 영화 '은밀하게 위대하게'의 네이버 영화평 중에서 평점 1점과 10점에 해당하는 문서를 각각 2,940개 5,555를 수집한 후 그림 1의 알고리즘을 적용하여 대표적인 긍정 영화평(평점 10)과 부정 영화평(평점 1)를 각각 50개씩 선정하였다. 요약 문서로 선정된 영화평들에 대한 특성분포 결과는 그림 3~5와 같다. 그림 3은 '스토리' 토픽에 대한 특성의 분포를 보여주며, a)와 b)는 각각 긍정과 부정 영화평에 대한 결과이다. 동일하게 그림 4와 5는 각각 '연출'과 '배우'에 대한 토픽들의 분석 결과이다. 각각에 대해서 '원본'은 전체 영화평 문서에서의 특성분포를 나타내며, β=0 ~ β=5는 그림 1의 알고리즘에서 라인 7의 β값을 적용했을 때 선택된 요약 문서에서의 특성분포를 나타낸다. 즉, β=0은 요약 문서를 선택할 때 마다 $r(f)$를 재조정하지 않는 경우이며, β값이 커질수록 재조정 비율이 높아지는 것을 의미한다.

우선 그림 3을 보면 긍정/부정 모두에 대해서 β=0일 경우 '원본'에 비해 요약결과에 '웹툰' 특성에 대한 내용이 지나차게 많은 것을 확인할 수 있으며, β값이 커질수록 점점 작아짐을 알 수 있다. 특히 긍정 영화평의 경우 β=1 또는 β=3일 경우 특성들의 분포가 원본과 매우 유사함을 확인할 수 있다. 따라서 그림 1의 알고리즘에서 요약 문서를 선택할 때 $r(f)$를 재조정하는 것이 매우 중요하다는 것을 알 수 있다. 이러한 현상은 그림 4와 5에서도 동일하게 확인할 수 있다. 다만 그림 5에서 긍정 영화평의 경우에는 β값에 큰 영향이 없었다. 하지만 그림 3~5의 결과에서 보듯이 본 논문이 제안한 요약 기법을 사용함으로써 '원본'과 유사한 특성분포를 갖는 요약 문서를 생성할 수 있다는 것이 입증되었고, β값을 조절하여 좀 더 정교한 요약이 가능함을 확인할 수 있었다.

VI. 결론

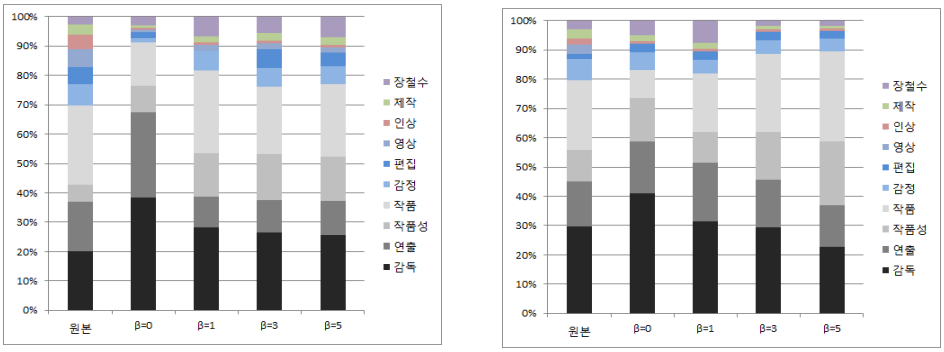
본 논문에서는 네이버 영화평과 같은 대량의 단문 오피니언 문서들을 대표하는 요약 문서들를 선정하는 방법을 제안하였다. 제안된 방법에서는 대상 객체의 특성들을 정의하고 특성 분포를 파악하여 이와 유사한 분포를 갖는 요약 문서들을 선정하였다. 요약 문서를 선정하는 기준으로는 주어진 토픽과의 연관성, 문서의 길이, 특성의 출현빈도 등을 고려하였고, 실험결과 제안된 기법이 원본과 유사한 특성 분포를 갖는 요약문서들을 생성한다는 것을 확인하였다. 본 논문이 제안한 기법은 영화평을 중심으로 설계되었으나, 상품평과 같은 단문의 오피니언 문서에 모두 적용 가능하며, 트위터와 갖는 SNS 문서에도 적용가능하다. 다만 SNS와 같은 문서들은 오피니언 문서들이 다른 성격의 문서들과 뒤섞여있어 이를 선별하는 작업이 우선시 되어야 하며, 이 문제를 해결하는 후속 연구를 진행할 계획이다.



(a) 긍정 영화평 (b) 부정 영화평

그림 3. '스토리' 토픽의 요약결과

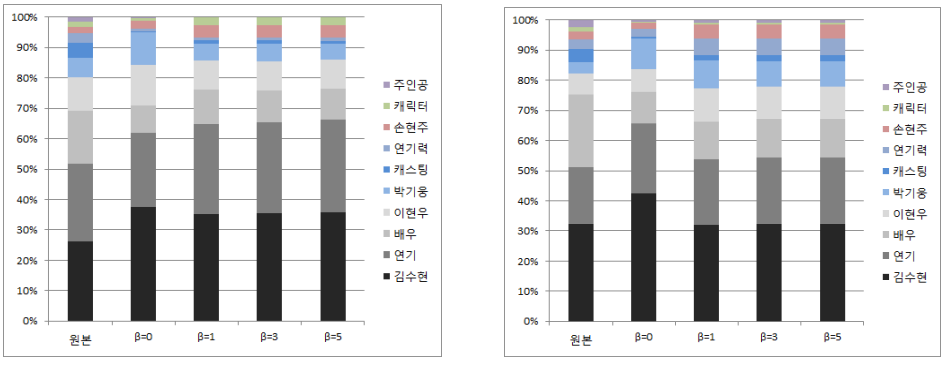
Fig. 3. Summary Results of Topic 'Story'



(a) 긍정 영화평 (b) 부정 영화평

그림 4. '연출' 토픽의 요약결과

Fig. 4. Summary Results of Topic 'Direction'



(a) 긍정 영화평 (b) 부정 영화평

그림 5. '배우' 토픽의 요약결과

Fig. 5. Summary Results of Topic 'Act'

References

[1] B. Liu, M. Hu, and J. Cheng, Opinion observer: analyzing and comparing opinions on the Web, Proceedings of the 14th international conference on WWW, pp. 10-14, 2005.

[2] C. Scaffdi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, Red Opal: Product-Feature Scoring from Reviews, Proceedings of the 8th ACM conference on Electronic commerce, pp. 11-15, 2007.

[3] Xiaowen Ding, and Bing Lui, The Utility of Linguistic Rules in Opinion Mining, Proceedings of SIGIR 2007, pp. 811-812, 2007.

[4] E. Courses, and T. Surveys, Using SentiWordNet for multilingual sentiment analysis, Proceedings of IEEE 24th International Conference on Data Engineering Workshop, ICDEW 2008, 2008.

[5] A. Popescu, and O. Extracting product features and opinions from reviews, Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp. 339-396, 2005.

[6] J. Liu, Y. Cao, C. Lin, Y. Huang, and M. Zhou, Low-Quality Product Review Detection in Opinion Summarization, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 334 - 342, 2007.

[7] A Pak, and P Paroubek, Twitter as a Corpus for Sentiment Analysis and Opinion Mining, Proceedings of The International Conference on Language Resources and Evaluation, pp. 1320-1326, 2010.

[8] R. Mihalcea and P. Tarau, TextRank: Bringing order into texts. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, 2004.

[9] X. Wan, TimedTextRank: Adding the Temporal Dimension to Multi-Document Summarization, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, pp. 867- 868, 2007.

[10] Y. Ouyang, W. Li and Q. Lu, An Integrated Multi-document Summarization Approach based on Word Hierarchical Representation, Proceedings of the ACL-IJCNLP Conference Short Papers, Suntec, Singapore, pp. 113 - 116, 2009.

[11] N. Garg, B. Favre, K. Reidhammer, D. Hakkani-Tuer, ClusterRank: A Graph Based Method for Meeting Summarization, Proceedings of Interspeech 2009: 10th Annual Conference Of The International Speech Communication Association, Vols 1-5 (ISBN: 978-1-61567-692-7), pp. 1507-1510, 2009.

[12] D. R. Radev, et al. Centroid-based summarization of multiple documents, Information Processing & Management, Vol. 40, No. 6, pp. 919-938. 2004.

[13] H. Zha, Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 113-120, 2002.

[14] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miler, Introduction to WordNet: An on-line lexical database, International Journal of Lexicography, pp. 235-244. 1990.

[15] G. Carenini and J. C. K. Cheng, Extractive vs. NLG-based abstractive summarization of evaluative text: The effect of corpus controversiality. In: Proceedings of the Fifth International Natural Language Generation Conference. Association for Computational Linguistics, p. 33-41, 2008.

[16] M. Litvak, and M. Last, Graph-based keyword extraction for single-document summarization, Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization. Association for Computational

Linguistics, 2008.

[17] http://en.wikipedia.org/wiki/PageRank

[18] F. Li, et al. Structure-aware review mining and summarization. Proceedings of the 23rd international conference on computational linguistics. Association for Computational Linguistics, pp. 653-661, 2010.

[19] G. Somprasertsri, and L. Pattarachai, Feature-Opinion in Online Customer Reviews for Opinion Summarization. J. of UCS, Vol. 16, No. 6, pp. 938-955, 2010.

[20] Y. Lu, and C. Zhai, N. Sundaresan, Rated aspect summarization of short comments, Proceedings of the 18th international conference on World wide web, pp. 131-140, 2009.

[21] X. Meng, et al. Entity-centric topic-oriented opinion summarization in twitter. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 379-387, 2012.

[22] H. KIM, et al. Comprehensive review of opinion summarization. Technical report, University of Illinois at Urbana-Champaign, 2011.

[23] M. Hu, and B. Liu, Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177, 2004.

[24] L. Zhuangm, L. Huang, F. Jing, and X. Zhu, Movie review mining and summarization, Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 43-50, 2006.

[25] J. Chang, and I. Kim, An Experimental Evaluation of Short Opinion Document Classification Using A Word Pattern Frequency, Journal of the Institute of Internet, Broadcasting and Communication, Vol. 12, No. 5, 2012.

[26] J. Shim, and H. C. Lee, The Development of Automatic Ontology Generation System Using Extended Search Keywords, Journal of the Korea Academia-Industrial cooperation Society, Vol. 11, No. 6, 2009.

저자 소개

장 재 영(정회원)

- 1992년: 서울대학교 계산통계학과 (이학사)
- 1994년: 서울대학교 계산통계학과 (이학석사)
- 1999년: 서울대학교 계산통계학과 (이학박사)
- 2000년 ~ 현재: 한성대학교 컴퓨터 공학과 교수

<주관심분야 : 데이터베이스, 정보검색, 데이터마이닝>

※ 이 논문은 2011년도 정부(교육부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임. (과제번호: NRF-2011-0022445),