사건중심 뉴스기사 자동요약을 위한 사건탐지 기법에 관한 연구*

A Study on an Effective Event Detection Method for Event-Focused News Summarization

정영미(Young-Mee Chung)**
김용광(Yong-Kwang Kim)***

초 록

이 연구에서는 사건중심 뉴스기사 요약문을 자동생성하기 위해 뉴스기사들을 SVM 분류기를 이용하여 사건 주제범주로 먼저 분류한 후, 각 주제범주 내에서 싱글패스 클러스터링 알고리즘을 통해 특정한 사건 관련 기사들을 탐지하는 기법을 제안하였다. 사건탐지 성능을 높이기 위해 고유명사에 가중치를 부여하고, 뉴스의 발생시간을 고려한 시간벌점함수를 제안하였다. 또한 일정 규모 이상의 클러스터를 분할하여 적절한 크기의 사건 클러스터를 생성하도록 수정된 싱글패스 알고리즘을 사용하였다. 이 연구에서 제안한 사건탐지 기법의 성능은 단순 싱글패스 클러스터링 기법에 비해 정확률, 재현율, F-척도에서 각각 37.1%, 0.1%, 35.4%의 성능 향상률을 보였고, 오보율과 탐지비용에서는 각각 74.7%, 11.3%의 향상률을 나타냈다.

ABSTRACT

This study investigates an event detection method with the aim of generating an event-focused news summary from a set of news articles on a certain event using a multi-document summarization technique. The event detection method first classifies news articles into the event related topic categories by employing a SVM classifier and then creates event clusters containing news articles on an event by a modified single pass clustering algorithm. The clustering algorithm applies a time penalty function as well as cluster partitioning to enhance the clustering performance. It was found that the event detection method proposed in this study showed a satisfactory performance in terms of both the F-measure and the detection cost.

키워드: 사건탐지, 사건중심 뉴스기사 자동요약, 지지벡터기, 싱글패스 알고리즘, 시간벌점함수
event detection, event-focused news summarization, support vector machine classifier, single pass algorithm, time penalty function

---

* 이 연구는 2006년도 연세대학교 학술연구비의 지원으로 이루어진 것임.
** 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)
*** 연세대학교 문헌정보학과 대학원(ykkim@yonsei.ac.kr)
■ 논문접수일자 : 2008년 11월 18일 ■ 최초심사일자 : 2008년 11월 20일 ■ 게재확정일자 : 2008년 12월 3일
■ 情報管理學會誌, 25(4): 227-243, 2008. 〔DOI:10.3743/KOSIM.2008.25.4.227〕

1. 서 론

오늘날 온라인으로 접근할 수 있는 정보의 양이 급증하면서, 이를 통제하고 조직하여 이용자의 정보 접근을 용이하도록 하는 다양한 기법들이 연구되고 있다. 특히 세계 각처에서 발생하는 사건(event)이나 이슈(issue)에 대한 관심이 높아지고, 보도되는 뉴스의 수가 급격히 증가함에 따라 개인이 이러한 뉴스를 효과적으로 검색하거나 브라우징 하는 데 도움을 줄 수 있는 도구가 필요하게 되었다.

1996년 미국 정부의 지원으로 시작된 TDT (Topic Detection and Tracking) 연구 프로젝트에서는 '사건' 관련 뉴스기사의 특성을 반영한 새로운 검색 기법들이 연구되었다. TDT 연구에서 정의한 '사건'이란 현실 세계에서 발생하는 구체적이고 실제적인 일을 지칭하는 것으로 정보검색에서 흔히 말하는 '주제'와는 구별된다(Allan 2002). 예를 들어, 주제적인 성격을 갖는 '항공기 추락 사고'는 시간이나 장소에 관계없이 항공기 추락과 관련된 모든 사고를 지칭하는 반면, 'F-15K 훈련 중 추락 사고'는 특정 시간과 장소에서 발생한 사고를 가리키며, 시간이 어느 정도 경과하면 이와 관련된 정보는 더 이상 생산되지 않는다. 따라서 뉴스의 발생시간이 사건을 탐지하는 데 중요한 요소가 되며, 또한 사건이 발생한 장소, 인명, 단체명 등의 고유명이 중요한 역할을 하게 된다.

TDT 연구의 사건탐지(event detection) 과제에서는 지속적으로 발생하는 뉴스기사를 대상으로 클러스터링 기법을 이용하여 같은 사건의 기사들을 모아 사건 클러스터를 생성한다. 이를 통해 이용자에게 동일한 사건에 대한 기사들을 함께 보여주거나 탐지된 일련의 기사들로부터 현재까지의 사건 개요를 보여 주는 하나의 종합적인 요약문을 작성하여 제시할 수 있다. 사건탐지 기법은 실시간으로 유입되는 뉴스를 대상으로 한 온라인 탐지(on-line detection)와 이전에 식별되지 않은 사건들을 대상으로 하는 소급적 탐지(retrospective detection)로 나뉜다. 사건탐지에서는 학습문헌을 사용하기 어렵기 때문에 주로 클러스터링 기법을 이용하며, 특히 온라인 사건탐지에서는 싱글패스(single pass) 클러스터링 알고리즘과 같이 재배치 작업이 없는 순차적 처리 알고리즘을 주로 사용한다(Yang et al. 1999; Leek et al. 2002).

그러나 단순한 싱글패스 알고리즘은 보통 성능이 좋지 않기 때문에 다양한 요소를 반영함으로써 성능을 높이려는 연구들이 시도되었다. 예컨대 사건의 시간 요소를 반영한 선형감소 가중치 함수를 사용하거나(Yang et al. 1999), 시간 벌점(time penalty)을 이용하여 클러스터 임계치를 조정하는 변형된 싱글패스 알고리즘을 사용하였다(Papka and Allan 1999). 또한 Leek 등(2002)은 확률 모형을 바탕으로 K-means 클러스터링 기법을 순차적인 알고리즘으로 변형하여 사용하였고, Dharanipragada 등(2002)은 2-계층 클러스터링(two-tiered clustering) 기법을 통해 성능을 향상시키고자 하였다.

이 연구는 실시간 뉴스기사 검색 시스템에서 사용할 수 있는 효과적인 사건 탐지 기법을 제안하고, 탐지된 일련의 뉴스기사들로부터 종합적인 사건 요약문을 작성하는 것을 목적으로 한다. 사건탐지의 성능을 높이기 위해 자동분류 기법을 통해 뉴스기사를 사건 주제범주로 먼저 분류한 후 각 주제범주에 속한 기사들에

대해 변형된 싱글패스 알고리즘을 적용하였다. 사건탐지 기법의 성능 평가에는 정보검색에서 일반적으로 사용되는 정확률, 재현율, F-척도와 사건탐지 척도로 흔히 사용되는 누락률, 오보율, 탐지비용을 사용하였다.

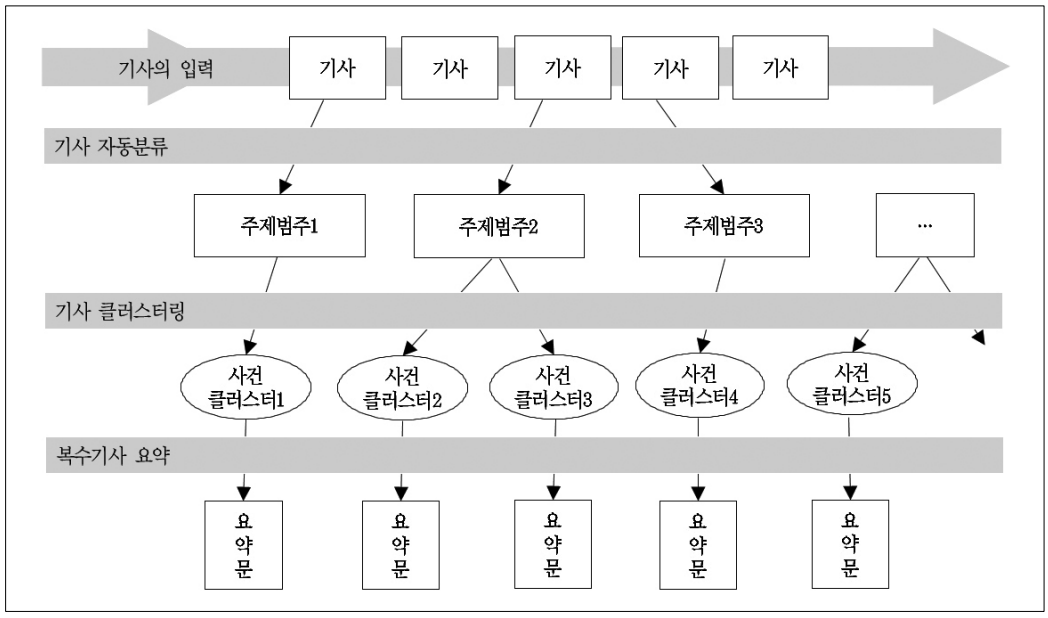
2. 실험 설계

2.1 실험 개요 및 실험문헌집단

이 연구에서 수행된 실험은 〈그림 1〉과 같이 크게 세 단계로 구성된다. 첫째, 온라인 환경에서의 사건탐지를 위해 같이 실시간으로 유입되는 기사들을 학습문헌을 이용하여 사건 주제범주로 분류한다. 둘째, 각 주제범주 내의 기사들을 군집화하여 사건 클러스터를 생성한다. 셋째, 생성된 사건 클러스터로부터 특정한 사건에 대한 요약문을 생성한다.

실험문헌집단은 한겨레신문과 동아일보의 종합, 북한, 경제, 사회, 특집, 국제/외신면으로부터 추출한 총 64,490개 뉴스기사로 구성하였다. 이 기사들에는 수작업으로 주제범주를 부여한 다음 사건탐지 실험을 위해 동일 사건을 보도하는 기사의 수가 충분한 사건들을 선정하였다.

먼저 뉴스기사의 주제범주 분류를 위해 한국언론연구원(1991)에서 발행한 '기사자료표준분류표'를 이용하였다. 이 분류표는 사건/사고에 해당하는 대주제 '500'대를 총 62개의 세부 주제로 분류하고 있는데, 이 연구에서는 이들 세부 주제 중 유사한 주제들을 통합하여 총 40개의 세부 주제범주를 〈표 1〉과 같이 설정하였다.



〈그림 1〉 사건탐지 및 자동요약 실험 개요

〈표 1〉 사건 주제범주

51 정치 사건
- 511 정치사건 일반
- 512 정치테러
- 513 반체제/반정부 운동
- 514 정치 비리/부정

52 군경 사건
- 521 군인, 경찰 사고, 부정/부패
- 522 군탈영
- 523 범죄자 탈옥/도망
- 524 군경 총기 사고

53 경제 사건
- 531 금융기관 사고
- 532 기업체 부도
- 533 외화유출
- 534 탈세
- 535 상거래 부정, 기업 비리/부정
- 536 밀조/밀매

54 사회 사건
- 540 파업/시위(기업, 학생 포함)
- 541 강도/살인
- 542 폭행/상해
- 543 절도/도난
- 544 유괴/납치/실종
- 545 성(性)범죄
- 546 보건 및 환경 범죄
- 547 사기(위조/변조, 공갈/협박, 횡령)
- 548 자살/사망

55 교육 사건
- 551 사학비리
  (부정 편/입학, 부정시험/입시부정 포함)
- 552 교수/교사 부정 및 비행
  (논문 표절 등 포함)
- 553 학생 비행

56 사고
- 561 교통사고
- 562 해상사고
- 563 항공사고
- 564 화재사고
- 565 폭발사고
- 566 기타사상사고
  (감전/쇼크, 우발사고, 안전사고, 등반사고, 산업재해 등)
- 567 보건사고/의료사고
  (전염병, 질병 발생)

57 자연재해
- 571 기상이변
- 572 풍수해
- 573 설해
- 574 가뭄
- 575 낙뢰
- 576 화산재해/지진

64,490개의 뉴스기사를 수작업 분류한 결과 2005년도 38,241건의 기사 중 3,807개, 2006년도 26,249개의 기사 중 2,127개의 기사가 사건 기사로 선정되었다. 이 중 2005년의 기사는 자동분류를 위한 학습문헌으로 사용하였고, 2006년 기사는 분류 검증을 위해 사용하였다. 또한 2006년의 기사들이 보도하는 사건들로부터 총 27개의 사건을 선정하였다. 따라서 사건탐지 실험에서는 2006년의 26,249개의 기사들로부터 27개의 사건과 관련된 기사들을 탐지해 내는 것이 목표가 된다. 사건탐지 실험을 위해 선정된 각 사건은 최소 5개부터 최대 261개의 기사들을 갖고 있으며, 각 사건별 기사 수와 사건 개요가 〈표 2〉에 나와 있다.

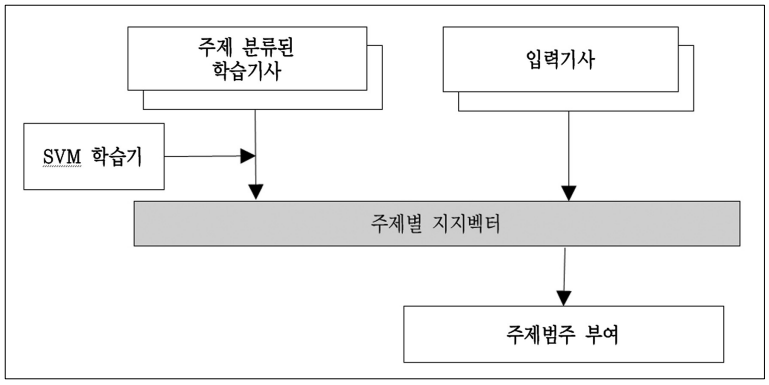
2.2 뉴스기사의 주제범주 자동분류

실시간으로 입력된 뉴스기사의 자동분류를

〈표 2〉 사건탐지 실험용 사건 관련 정보

사건 번호	기사 수	사건 내용	사건 번호	기사 수	사건 내용
E01	8	전국 연쇄 성폭행 사건	E15	5	전 청와대 행정관의 아내 살해 사건
E02	5	부자 초등생(아동) 성추행 살해 사건	E16	8	네팔 독재 국왕 하야 요구 국민 시위
E03	19	철도노조 총파업	E17	8	비디오 아티스트 백남준 사망
E04	7	놀이동산(롯데월드) 안전사고	E18	7	나이지리아 무장단체 한국인 납치
E05	5	서울 연쇄 성폭행범(마포 발바리) 검거	E19	14	소말리아 해적의 한국어선 피랍/납치
E06	31	중남부 폭우 재해	E20	10	필리핀 국민 반정부 시위
E07	5	영국 런던 항공기 테러	E21	5	F-16 전투기 군산 앞바다 추락
E08	5	미국 카트리나 발생	E22	6	에어쇼 전투기 추락 및 조종사 사망
E09	7	전세계 조류 독감	E23	15	F-15K 야간 훈련비행 중 동해상 추락
E10	7	브라질 폭력조직의 경찰서 습격	E24	7	론스타 외환은행 매각 및 외환법 위반
E11	7	이종욱 WHO 사무총장 사망/별세	E25	6	휴대폰 제조업체 VK 부도
E12	29	서래마을 프랑스인 영아 유기	E26	184	사행성 게임 사업 관련 비리
E13	70	한나라당 박근혜 대표 피습	E27	261	황우석 논문조작 의혹
E14	5	미군 이라크민간인 성폭행/살인 사건			

위해 이 연구에서는 SVM(Support Vector Machine) 분류기를 사용하였다. SVM 분류기는 긍정예제와 부정예제를 가장 잘 분리하는 결정면(hyperplane)을 찾아내는 기법으로, 분류기의 학습 목표는 최적의 결정면 근처의 점들인 지지벡터(support vector)를 찾는 자동분류 기법으로 여러 분류기 중 가장 좋은 성능을 나타내는 것으로 알려져 있다(정영미 2005). 이 연구에서는 분류기 구현을 위해 Chang and Lin(2001)이 공개한 LibSVM을 사용하였다. LibSVM은 C언어와 자바를 비롯한 다양한 프로그래밍 언어에서 활용할 수 있도록 라이브러리 형태로 공개된 프로그램으로 이 실험에서는 Python용 라이브러리를 이용하였다.



〈그림 2〉 주제범주 자동분류 개요

입력된 기사를 해당되는 사건 주제범주로 분류하는 과정은 〈그림 2〉와 같이 학습과 실제 주제범주 부여의 두 단계로 구성된다. 첫 번째 단계에서는 수작업으로 주제범주가 부여된 학습용 뉴스기사들로부터 SVM 학습기를 통해 주제별 지지벡터를 생성한다. 두 번째 단계에서 분류 대상 기사가 입력되면 주제범주별 지지벡터를 이용하여 각 주제범주에 대한 분류확률을 계산하여, 가장 높은 분류확률을 갖는 주제범주로 기사를 분류하게 된다.

모든 기사에 대해 "21세기 세종계획"에서 개발한 〈지능형형태소분석기 버전 2.0〉을 이용하여 형태소를 분석하였고, 각 기사로부터 일반명사 및 고유명사를 색인어로 추출하였다. 주제범주 분류 실험에서는 학습기사 및 입력기사의 분류자질로 사건 클러스터링 과정에서와는 달리 일반명사만을 사용하였다. 사전 실험에서도 고유명사를 제외한 일반명사만을 자질로 사용하였을 때 분류 성능이 높게 나타났다.

색인어로 선정된 각 용어에는 아래와 같은 공식의 Okapi TF 가중치를 부여하였다.

$$Okapi\ TF = \frac{(k_1+1)tf}{k_1((1-b)+b \cdot \frac{doc.length}{avg\ doc.length})+tf}$$

위 공식에서 tf는 기사 내 용어빈도, $doc.length$는 해당 기사의 길이, 그리고 $avg\ doc.length$는 모든 기사의 평균 길이를 의미한다. k_1과 b는 상수로 각각 TREC-6 실험에서 사용한 1.2와 0.75를 사용하였다(Walker et al. 1998).

앞에서 언급했듯이 38,241개의 전체 학습기사 가운데 사건 주제범주에 해당하는 기사는 3,807개로서 전체의 약 10%를 차지한다. 전체 기사를 모두 학습에 사용할 경우 실제 입력기사 분류 시 사건 주제에 해당하는 기사가 잘못 분류될 가능성이 크다. 사건에 해당하는 기사만을 학습에 이용하면 학습의 효율이 좋기 때문에 온라인 사건탐지 환경에서 이러한 학습방법이 특히 적합하다. 사건에 해당하는 기사만을 학습기사로 하여 실제 입력된 기사를 분류해 본 결과, 사건 주제에 해당하는 기사가 아닐 경우 각 사건 주제범주에 대한 분류 확률이 현저하게 낮은 것으로 나타났다. 따라서 주제범주 분류 실험 시, 입력기사의 각 주제범주별 분류확률을 계산한 다음 가장 큰 분류확률 값이 임계치 이하인 경우 해당 입력기사를 사건 주제범주로 분류하지 않음으로써 입력기사 중 잡음 데이터, 즉 사건에 해당하지 않는 기사를 제거하였다.

한편, 사건 주제범주에 정확하게 분류되지 않은 기사가 많을 경우, 사건 클러스터 생성 시 관련 기사가 배제되므로 클러스터링 성능에 큰 영향을 미치게 된다. 따라서 주제범주 분류에서는 재현율을 높이는 것이 중요하다. SVM 분류기에서 적합문헌이 각 범주에 속할 확률을 추정할 수 있는데, 분류확률이 1위인 최적합 범주의 추정 확률이 0.7 미만인 경우 2위나 3위의 범주까지 적합한 범주로 복수 분류함으로써 재현율을 높일 수 있다. 실제 분류 실험 결과 2개의 범주에 분류할 경우 재현율이 0.63에서 0.74로 높아지고, 3개의 범주에 할당할 경우 0.80까지 향상되었다. 따라서 실제 분류 실험에서는 3개까지의 복수 범주에 뉴스기사를 분류하였다.

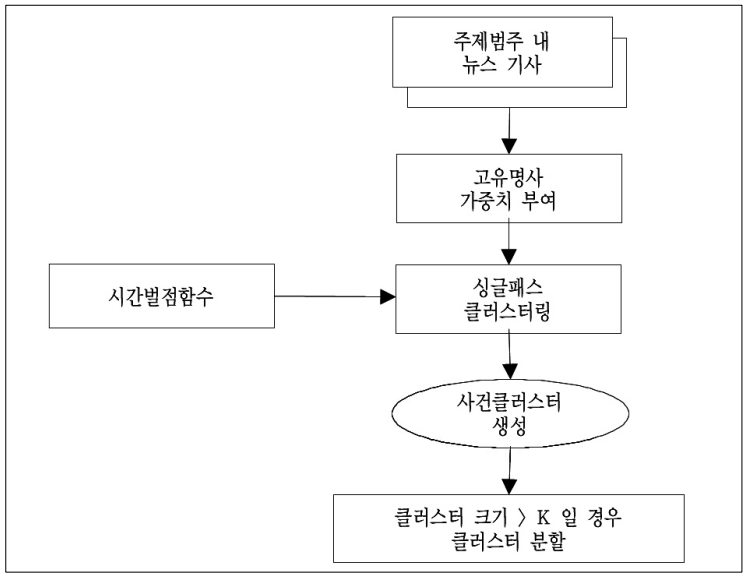
2.3 사건 클러스터 생성

일단 뉴스기사의 주제범주가 부여되면 해당 주제범주 내 기사들로부터 사건 클러스터를 생성한다. 온라인 환경의 사건탐지 실험에서는 입력기사가 실시간으로 처리되어야 하므로 주로 싱글패스 클러스터링 알고리즘을 사용한다. 싱글패스 알고리즘에서는 처음 입력된 문헌이 첫 클러스터가 되며, 다음에 입력된 문헌과 기존 클러스터와의 유사도를 산출하여 임계치를 넘으면 유사도 값이 최대인 클러스터에 배정하고, 그렇지 않으면 새로운 클러스터가 생성된다.

일반적으로 싱글패스 알고리즘은 단순하고 컴퓨터 처리 시간 면에서 효율적이지만, 클러스터링 성능이 비교적 좋지 않은 것으로 알려져 있다. 이 연구에서는 이러한 싱글패스 알고리즘의 성능을 높이기 위해 〈그림 3〉에서와 같이 사건 클러스터 생성 과정에 고유명사 가중치 부여, 시간벌점함수의 적용, 클러스터 분할 등 세 가지 방법을 도입하였다.

2.3.1 고유명사 가중치 부여

특정한 사건을 식별하는 데 있어서 지명, 인명, 단체명 등 고유명사의 역할이 매우 중요하므로 이러한 고유명사에는 일반명사보다 큰 가중치를 부여하였다. 예를 들어 산불이 발생했을 때, 산불과 관련된 일반명사들은 해당 기사가 산불과 관련된 기사, 즉 산불이라는 주제범주의 기사임을 판정하는 데에는 도움이 되지만 특정한 사건을 식별하는 데 있어서는 산불이 발생한 지역의 이름이 중요한 역할을 한다. 외국 연구에서는 이미 개발된 고유명 인식기를 사건 클러스터를 생성에 사용하는데, 한글의 경우 공개된 고유명 인식기가 없을 뿐만 아니라 영어에 비해 한글 고유명 인식 자체가 어렵다.



〈그림 3〉 사건 클러스터 생성 개요

따라서 이 연구에서는 고유명사를 최대한 식별하고 이에 가중치를 부여함으로써 클러스터링 성능을 높이고자 하였다. 이 연구에서 사용한 〈지능형형태소분석기〉는 국가명, 도시명, 한국인명 등은 비교적 잘 분석하지만, '론스타', '카트리나' 등의 외래어로 된 고유명사를 제대로 분석하지 못한다는 단점이 있다. 이와 같은 외래어는 조사와 분리되지 않은 채 주로 '/NF'라는 '명사추정범주'로 분석되는 경우가 많았다. 예를 들어 '론스타가 외환은행을 매각하는...'이라는 문장의 '론스타가'는 조사 '가' 앞의 '론스타'가 고유명사로 식별되지 못하고 명사추정범주로 분석된다. 이 연구에서는 이러한 경우 조사 '가'를 분리하고 '론스타'를 고유명사로 추출하였다. 조사를 분리하기 위해 어절의 뒷부분에서 가장 긴 조사가 일치할 경우 분리하는 '최장조사일치' 기법을 이용하였다.

2.3.2 시간벌점함수 사용

사건 클러스터링 실험에서 뉴스기사의 발생 시간 요소를 고려한 시간벌점함수를 사용하여 싱글패스 알고리즘의 성능을 높이고자 하였다. 기사의 발생 시간을 고려한 기존의 연구에서는 단순히 시간창(time window)의 크기를 제한하거나 또는 시간 경과에 따라 단순히 선형적으로 기사와 사건 클러스터 간 유사도를 낮추어 주는 방법을 사용하거나(Yang et al. 1999) 사건 클러스터의 임계치를 높여 주는 방법(Papka and Allan 1999) 등을 사용하였다. 그러나 이러한 방법들은 사건별 기사 발생 기간의 분포를 고려하고 있지 않고 있다. 실제로 이 연구에서의 특정한 사건 관련 기사들의 발생 기간은 최소 1일부터 최대 329일로 매우 다양하고, 또한 사건별로 특정 기간에 매우 집중적으로 기사가 발생하거나 혹은 긴 기간에 걸쳐 띄엄띄엄 기사가 발생하는 등 기사의 발생 분포도 다양한 것으로 나타났다. 따라서 아래와 같은 시간벌점함수 공식을 이용하여 입력기사와 사건 클러스터 벡터와의 유사도 값을 조정하였다.

$$if \quad tpf(d_i,c) > 0 :$$
$$sim(d_i,c) = tpf(d_i,c) \times sim(d_i,c)$$
$$\left(tpf(d_i,c) = 1 - \frac{td_i}{\beta \times mean(td_1 \sim td_{i-1})} \right)$$
$$if \quad tpf(d_i,c) \le 0 : \; sim(d_i,c) = 0$$

위 공식에서 $sim(d_i,c)$는 입력기사와 기존에 생성된 사건 클러스터 c와의 유사도를 의미하고 $tpf(d_i,c)$는 시간벌점함수를 의미한다. 시간벌점함수가 0보다 작을 경우 유사도 값이 0이 되고, 0보다 큰 경우 시간벌점함수에 따라 유사도 값이 조정된다. 시간벌점함수에서 td_i는 클러스터의 최근 기사 d_{i-1}과 d_i와의 시간 차이(time difference)을 의미하고, β는 상수이다. $mean(td_1 \sim td_{i-1})$은 클러스터 내 기사들의 시간 차이의 평균을 의미한다. 공식에서 td_i가 클수록 즉, 입력된 기사와 사건 클러스터 내 최근 기사와의 시간 차이가 클수록 tpf의 값이 작아지면서 입력기사와 클러스터와의 유사도가 감소된다. 한편, $mean(td_1 \sim td_{i-1})$이 클수록 즉, 특정 클러스터 내의 기사들의 기간 차이의 평균이 높을수록 tpf의 값이 커지므로 유사도가 감소하는 정도가 작아지게 된다.

즉, 하나의 사건이 특정 기간에 집중적으로 발생할 경우 $mean(td_1 \sim td_{i-1})$이 작아지

고, 입력된 기사의 td_i의 값이 클 경우 입력기사와 사건 클러스터와의 유사도 값이 감소되는 정도가 커지게 되므로 기사가 해당 사건에 분류될 확률이 작아져서 해당 기사가 누락될 가능성이 커지게 된다. 반대로 클러스터 내 기사들의 발생 시간 차이의 평균이 클 경우 입력된 기사의 td_i 값이 크더라도 유사도 값이 감소되는 정도가 작으므로 입력기사가 해당 사건 클러스터에 포함될 확률이 커진다. 또한 상수 β는 td_i와 $mean(td_1 \sim td_{i-1})$에 따라 조정되는 유사도의 정도에 영향을 미친다. β 값이 매우 큰 경우 td_i와 $mean(td_1 \sim td_{i-1})$의 값에 상관없이 시간벌점함수의 값이 1에 가깝기 때문에 사건과 관계없는 기사가 입력되더라도 유사도가 임계치 이상이면 해당 사건에 포함될 가능성이 커지게 된다. 반면에 너무 작을 경우 시간벌점함수 값이 0에 가깝게 되거나 혹은 음의 값을 가지므로 입력기사가 누락될 가능성이 커지게 된다. 이 실험에서는 β 값을 0(시간벌점함수를 사용하지 않을 경우)에서 70까지 변화시켜 실험하여 최적의 값을 경험적으로 찾아내고자 하였다.

2.3.3 사건 클러스터 분할

클러스터링 결과 생성된 클러스터의 크기가 너무 크거나 작을 수 있다는 싱글패스 알고리즘의 단점을 보완하기 위해 일정 규모 이상의 클러스터를 분할함으로써 클러스터링 성능을 향상시키고자 하였다. 실제로 특정한 사건 관련 기사의 수는 10개 미만이지만 단순한 싱글패스 알고리즘을 적용하여 생성한 사건 클러스터들 중에는 포함된 기사 수가 200개 정도인 매우 큰 클러스터도 있었다. 클러스터 크기를 조정하기 위해 클러스터 내 기사 수가 K에 도달할 경우 클러스터 임계치를 일정 수치($\Delta\theta$) 증가시킨 다음 다시 싱글패스 알고리즘을 적용하여 클러스터를 분할하였다. 여기서 K와 $\Delta\theta$는 경험적으로 최적의 값을 찾아내어 사용하였다.

2.4 사건 탐지 성능 평가 척도

사건탐지 성능을 평가하기 위해 기존 TDT 연구의 성능 평가 척도인 누락률, 오보율, 탐지비용과 정보검색 연구에서 일반적으로 사용되는 정확률, 재현율, F-척도를 이용하였다. 누락률은 사건 클러스터에 포함되지 않은 적합기사의 비율로 누락된 정보의 양을 나타내고, 오보율은 사건 클러스터에 포함된 부적합기사의 비율로 부적합기사가 적합기사로 잘못 처리된 정도를 측정한다. 대표적인 사건탐지 척도인 탐지비용은 누락률과 오보율을 통합한 단일가 척도로 아래와 같은 공식에 의해 산출된다.

$$C_{det} = \alpha_1 \frac{b}{n} + \alpha_2 \frac{c}{n}$$

위 공식에서 b는 적합한 기사 중 검색되지 않은 기사의 수를, c는 검색된 부적합한 기사의 수를 각각 의미하며, n은 모든 기사의 수이다. 또한 α_1과 α_2는 상수로, TDT 표준 평가에서는 각각 0.1과 1을 사용하였다. 이는 적합기사가 누락되는 비율이 부적합기사가 적합기사로 잘못 판별되는 비율보다 중요하게 고려하여 시스템 성능을 평가하는 것을 의미한다(Yang et al. 2000). 누락률, 오보율, 탐지비용은 모두 값이 작을수록 좋은 성능을 나타낸다.

반면 재현율은 모든 적합한 기사 중 사건 클

러스터에 포함된 적합기사의 비율을, 정확률은 사건 클러스터에 포함된 기사 중 적합기사의 비율을 나타내는 척도로 그 값이 클수록 좋은 성능을 나타낸다. F-척도는 정확률과 재현율을 복합적으로 반영하는 단일가 척도로 TDT 연구에서도 많이 사용되고 있으며, 이 연구에서는 재현율과 정확률에 똑같은 비중을 준 F1척도를 사용하였다(정영미 2005). 최종적인 사건 탐지 성능은 각 사건에 대해 산출한 각 평가 척도 값의 평균을 산출하여 측정하였으며, 이를 위해 매크로 평가(macro evaluation) 기법을 사용하였다.

3. 실험 결과 분석

3.1 고유명사 가중치 부여에 따른 사건 탐지 성능

〈그림 4〉는 고유명사에 부여할 최적의 가중치 값을 찾아내기 위해 가중치 값을 달리 하여 단순 싱글패스 알고리즘을 적용한 결과를 보여준다. 그림의 가중치 값은 일반명사의 중요도를 1로 보았을 때 고유명사에 부여되는 중요도를 나타낸다. 실제 실험에서는 고유명사의 중요도를 1로 했을 때 일반명사에 0.5~1.0까지 0.1씩의 차이를 두고 중요도를 부여하였다. 즉 일반명사의 중요도를 0.5로 할 경우 원래의 OKAPI TF 값의 절반 값이 용어 가중치로 사용되는 것이다. 실험 결과 고유명사만을 클러스터링 자질로 사용했을 경우 일반명사와 고유명사 모두 가중치 없이 분류자질로 사용하였을 경우에 비해 정확률과 F 값이 낮은 것으로 나타났다. 고유명사에 일반명사의 약 1.43배의 가중치를 주었을 때 정확률, 재현율, F-척도에서 모두 가장 좋은 성능을 보였다. 따라서 이하 모든 실험에서 일반명사에는 1, 고유명사에는 1.43배의 용어 가중치를 부여하였다.

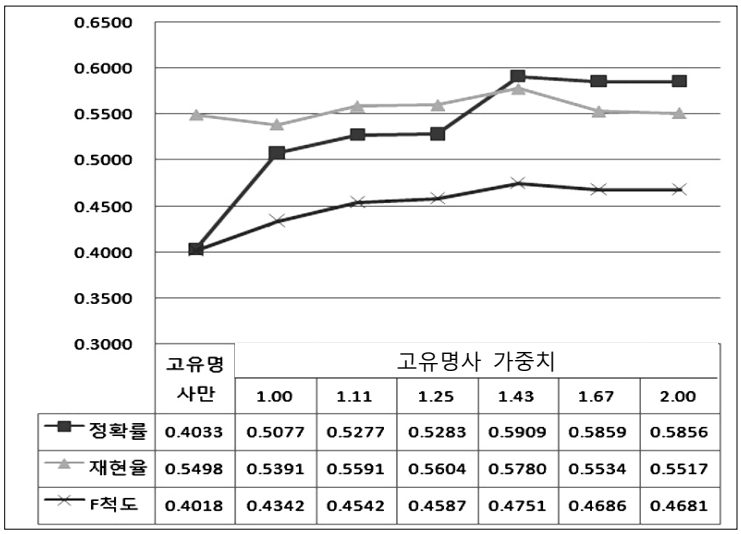
3.2 주제범주 활용에 따른 사건 탐지 성능

〈표 3〉은 단순 싱글패스 알고리즘만을 이용하여 사건 클러스터를 생성한 결과(S)와 뉴스기사를 사건 주제범주에 분류한 후 싱글패스 알고리즘을 적용한 실험 결과(C+S)를 보여준다. 두 실험 모두 클러스터 임계치가 0.31이었을 때 가장 좋은 성능을 보였다.

실험 결과, 베이스라인 시스템이 되는 S에 비해 C+S가 정확률, 재현율, F-척도에서 각각 약 7%, 29%, 21.7% 성능이 향상되었고, 누락률과 탐지비용에서는 각각 38.4%, 17.6% 정도 성능이 향상되었으나, 오보율에서만 성능이 다소 저하된 것으로 나타났다.

3.3 시간벌점함수에 따른 사건 탐지 성능

클러스터링 실험에서는 특정한 사건 클러스터에 속하는 기사들의 발생시간 분포를 고려한 시간벌점함수를 사용하여 기사와 사건 클러스터 간 유사도를 조정하였다. 〈그림 5〉는 단순 싱글패스 알고리즘에서 시간벌점함수만을 적용한 결과로서 시간벌점함수의 파라미터 β 값에 따른 성능의 변화를 보여 준다. 실험 결과 F 값을 기준으로 삼을 경우, β 값이 20 정도일

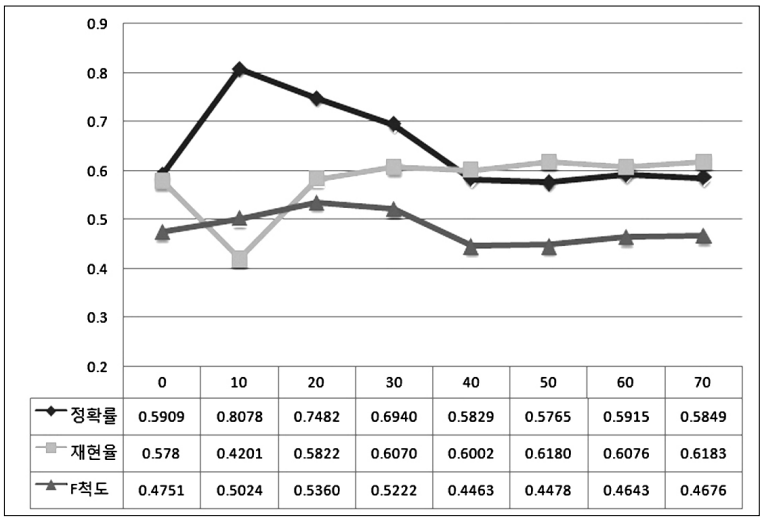


고유명사만	고유명사 가중치					
	1.00	1.11	1.25	1.43	1.67	2.00
정확률	0.4033	0.5077	0.5277	0.5283	0.5909	0.5859	0.5856
재현율	0.5498	0.5391	0.5591	0.5604	0.5780	0.5534	0.5517
F척도	0.4018	0.4342	0.4542	0.4587	0.4751	0.4686	0.4681

〈그림 4〉 고유명사 가중치에 따른 성능 변화

〈표 3〉 주제범주 활용에 따른 사건 탐지 성능 변화

	정확률	재현율	F척도	누락률	오보율	탐지비용
S	0.5909	0.5780	0.4751	0.4220	0.0186	0.0151
C+S	0.6323	0.7400	0.5780	0.2600	0.0197	0.0124
성능 차이	+0.0414	+0.1620	+0.1029	-0.1619	+0.0011	-0.0027
성능 향상률	+7.0%	+28.0%	+21.7%	+38.4%	-5.8%	+17.6%



	0	10	20	30	40	50	60	70
정확률	0.5909	0.8078	0.7482	0.6940	0.5829	0.5765	0.5915	0.5849
재현율	0.578	0.4201	0.5822	0.6070	0.6002	0.6180	0.6076	0.6183
F척도	0.4751	0.5024	0.5360	0.5222	0.4463	0.4478	0.4643	0.4676

〈그림 5〉 시간벌점 함수의 β에 따른 성능 변화

때 성능이 가장 높은 것으로 나타났다.

〈표 4〉는 시간벌점함수의 적용이 사건탐지 성능에 미치는 영향을 나타낸다. β 값을 20으로 한 시간벌점함수를 적용했을 때(S+TP) 정확률이나 F-척도는 각각 26.6%와 12.8%에 달하는 성능 향상률을 보였지만 재현율과 누락율에서는 미미한 향상률을 보였고, 오보율이나 탐지비용 측면에서는 오히려 성능이 떨어진 것으로 나타났다.

이와 같은 경향은 주제범주 분류 후 시간벌점함수를 적용했을 경우(C+S+TP)에서도 비슷하게 나타난다. 〈표 5〉와 같이 시간벌점함수를 적용했을 경우 정확률, F-척도, 오보율에서는 성능이 향상된 반면 재현율 및 누락률, 탐지비용에서는 오히려 성능이 감소하였다. 실험결과 시간벌점함수는 특히 사건탐지의 정확성을 높이는 데 효과가 있는 것으로 나타났다. 클러스터링 실험 결과 측정한 정확률이나 재현율 값 평균에 비해 F 값 평균이 낮게 나타난 경우가 많은 것은 매우 낮은 정확률이나 재현율을 갖는 사건들이 F 값을 낮추는 결과를 가져왔기 때문이다.

3.4 클러스터 분할에 따른 사건 탐지 성능

이 연구에서는 생성된 클러스터의 크기를 조정하기 위하여 사건 클러스터 내 기사 수가 K개 이상인 경우 클러스터 임계치를 일정 수치($\Delta\theta$) 증가시켜 클러스터를 분할하였다. 사전실험 결과 K는 100, $\Delta\theta$는 0.03일 경우 가장 좋은 성능을 보였기 때문에 모든 실험에서 이 값을 적용하였다.

〈표 6〉에서는 단순 싱글패스 알고리즘만 사용한 경우(S)와 클러스터 분할을 적용한 경우(S+CP)의 성능을 비교하였다. 클러스터 분할을 적용한 실험에서 정확률이 눈에 띠게 증가하였으나 재현율은 약간 감소하였고, F-척도도 증가하였다. 그러나 누락률과 탐지비용 측면에서의 성능은 약간 감소한 것으로 나타났

〈표 4〉 시간벌점 함수에 따른 사건 탐지 성능(단순 싱글패스 적용)

	정확률	재현율	F척도	누락률	오보율	탐지비용
S	0.5909	0.5780	0.4751	0.4220	0.0186	0.0151
S+TP	0.7482	0.5822	0.5360	0.4178	0.0256	0.0169
성능 차이	+0.1573	+0.0042	+0.0609	-0.0042	+0.0070	+0.0018
성능 향상률	+26.6%	+0.7%	+12.8%	+1.0%	-37.9%	-11.9%

〈표 5〉 시간벌점 함수에 따른 사건 탐지 성능(주제범주 할당 후 싱글패스 적용)

	정확률	재현율	F척도	누락률	오보율	탐지비용
C+S	0.6323	0.7400	0.5780	0.2600	0.0197	0.0124
C+S+TP	0.8577	0.5594	0.6357	0.4406	0.0047	0.0134
성능 차이	+0.2254	-0.1806	+0.0577	+0.1806	-0.0150	+0.0010
성능 향상률	+35.7%	-24.4%	+10.0%	-69.5%	+76.1%	-8.1%

〈표 6〉 클러스터 분할에 따른 사건 탐지 성능(단순 싱글패스 적용)

	정확률	재현율	F척도	누락률	오보율	탐지비용
S	0.5909	0.5780	0.4751	0.4220	0.0186	0.0151
S+CP	0.7349	0.5531	0.5302	0.4280	0.0167	0.0156
성능 차이	+0.1440	-0.0249	+0.0551	+0.0060	-0.0019	+0.0005
성능 향상률	+24.4%	-4.3%	+11.6%	-1.4%	+10.2%	-3.3%

다. 〈표 7〉과 같이 주제범주 분류 후 싱글패스를 적용한 시스템에서 클러스터 분할을 한 경우(C+S+CP)에도 유사한 경향이 나타났으며, 재현율과 누락률 측면에서의 성능이 〈표 6〉에 비해 크게 저하되었다. 그러나 베이스라인 시스템인 S에 비해서는 정확률, 재현율 F척도는 각각 37.1%, 0.1%, 35.4% 증가하였고, 오보율과 탐지비용에서의 성능도 각각 74.7%, 11.3% 증가하여 전반적으로 성능이 향상된 것으로 나타났다.

3.5 사건 탐지 기법 성능의 종합 평가

〈표 8〉은 이 연구에서 수행한 모든 사건탐지 실험 결과를 종합한 것이다. 표에서 제일 하단의 C+S+TP+CP가 이 연구에서 제안한 사건탐지 기법을 평가한 결과로서 정도의 차이는 있지만 베이스라인 시스템에 비해 모든 척도에서 좋은 성능을 나타냈다. 시간벌점함수(TP)와 클러스터 분할(CP)을 적용했을 경우 특히 정확률의 증가는 두드러지게 나타났고, 재현율은 다소 감소하는 경향이 나타났다. 그러나 사

〈표 7〉 클러스터 분할에 따른 사건 탐지 성능(주제범주 할당 후 싱글패스 적용)

	정확률	재현율	F척도	누락률	오보율	탐지비용
C+S	0.6323	0.7400	0.5780	0.2600	0.0309	0.0124
C+S+CP	0.8103	0.5844	0.6431	0.4406	0.0047	0.0134
성능 차이	+0.1780	-0.1556	+0.0651	+0.1806	-0.0262	+0.0010
성능 향상률	+28.2%	-21.0%	+11.3%	-69.5%	+84.8%	-8.1%

〈표 8〉 전체 사건 탐지 성능

		정확률	재현율	F척도	누락률	오보율	탐지비용
클러스터링	S	0.5909	0.5780	0.4751	0.4220	0.0186	0.0151
	S+TP	0.6211	0.5838	0.5143	0.4156	0.0122	0.0157
	S+CP	0.7349	0.5531	0.5302	0.4280	0.0167	0.0156
	S+TP+CP	0.7484	0.5668	0.5592	0.4158	0.0224	0.0154
분류 + 클러스터링	C+S	0.6323	0.7400	0.5780	0.2600	0.0309	0.0124
	C+S+TP	0.8577	0.5594	0.6357	0.4406	0.0047	0.0134
	C+S+CP	0.8103	0.5844	0.6431	0.4156	0.0211	0.0121
	C+S+TP+CP	0.8156	0.6533	0.6619	0.3467	0.0145	0.0113

건탐지의 대표적 척도인 F-척도와 탐지비용에서는 가장 좋은 성능을 보였다.

3.6 사건 요약문 작성

3.6.1 자동요약 알고리즘

사건탐지 기법에 의해 생성된 사건 클러스터로부터 하나의 사건 요약문을 작성하기 위해 뉴스 기사를 대상 복수문헌 요약 시스템인 MEAD 시스템(Radev et al. 2004)의 알고리즘을 사용하였다. MEAD 알고리즘에서는 우선 생성된 기사들의 클러스터의 센트로이드를 생성한 후, 각 기사를 구성하는 모든 문장의 중요도 점수를 산출한 후 압축률에 따라 상위의 몇 문장을 나열하여 요약문을 작성한다. 문장 i의 중요도 점수를 산출하기 위해 총 네 개의 요소로 이루어진 아래와 같은 공식을 사용한다.

$$SCORE(s_i) = w_c C_i + w_p P_i + w_f F_i - w_R R_s$$

위 식에서 w_c와 w_p, w_f는 최종점수를 구성하는 각 요소에 중요도를 나타내는 파라미터로서 MEAD 시스템에서 가장 좋은 성능을 보인 값은 각각 1, 2, 1이었다(Radev et al. 2004). C_i는 문장과 클러스터 센트로이드와의 유사도, P_i는 문장의 위치점수, F_i는 첫 문장과의 유사도를 나타낸다. 또한 R_s는 문장의 중복도를 나타내는 요소로 두 문장에서 중복되는 용어의 수에 2를 곱한 값을 두 문장의 용어 수로 나눈 값이며, 다른 모든 문장과의 중복도 값 중에서 최대값을 사용한다.

문장의 위치점수를 나타내는 P_i는 기사 내 문장의 수가 n 일 경우 아래와 같은 공식에 의해 산출한다. 이 식에서 $C_{\max}$는 C_i의 최대값을 의미한다.

$$P_i = \frac{(n-i+1)}{n} * C_{\max}$$

3.6.2 요약문 생성 결과

요약문 생성 실험에서는 가장 좋은 성능을 나타낸 C+S+TP+CP 시스템에서 탐지한 사건을 대상으로 앞에서 기술한 자동요약 기법을 이용하여 압축률 10%의 요약문을 생성하였다. 아래 제시한 요약문은 사건 번호 E11의 '이종욱 WHO 사무총장 사망' 사건 기사들로부터 생성한 요약문이다. E11 사건 탐지 결과 총 7개의 기사가 사건 클러스터에 포함되었는데 이 가운데 5개의 기사가 적합한 기사였다. 그러나 이 연구에서 사용한 자동요약 기법은 사건 클러스터로부터 주요 문장만을 추출하기 때문에 부적합 기사들의 문장은 요약문에 포함되지 않았다. 각 요약문장 [ ] 안의 숫자는 각각 기사의 번호와 기사 내 문장의 번호를 의미한다.

> 한국인 최초의 유엔 기구 수장인 이종욱(사진) 세계보건기구(WHO) 사무총장이 20일 오후 갑자기 쓰러져 병원으로 긴급 이송됐다. [37462 : 1]
>
> 이종욱(61, 사진) 세계보건기구(WHO) 사무총장이 22일 오전(현지시각) 스위스 제네바에서 숨졌다. [37590 : 1] 세계보건기구 사무국은 이날 오전 제59차 총회 개막에 앞서 "이 총장이 20일 집무 중 뇌출혈로 쓰러져 제네바 칸토날 병원에서 수술을 받은 뒤 숨졌다"고 밝혔다. [37590 : 2] 신영수 서울대 의대 교수(세계보건기구 사무총장 특별자문관)는 "조문을 위해 제

네바로 곧 출국할 예정"이라고 밝혔다. [37590 : 4] 이 총장은 세계보건기구 총회 개막에 앞서 회의 준비를 하던 중 20일 집무실에서 갑자기 쓰러져 구급차로 제네바 칸토날병원으로 이송됐다. [37590 : 5]

22일 오전(현지시각) 스위스 제네바에서 숨진 이종욱(61) 세계보건기구(WHO) 사무총장은 2003년 1월 한국인으로서는 처음으로 유엔 산하 전문기구의 수장에 뽑혀 유명해졌다. [37623 : 1] 이 총장은 1983년 세계보건기구 남태평양지역 피지에서 서태평양지역 사무처 한센병 자문관으로 일하면서 이 기구와 인연을 맺었다. [37623 : 2]

고 이종욱 세계보건기구(WHO) 사무총장이 국립묘지에 안장된다. [37689 : 1] 국가보훈처는 23일 국립묘지 안장대상 심의위원회를 열어 이종욱 사무총장이 국가사회 발전에 기여한 공로를 기려 국립묘지 안장 대상자로 결정했다고 밝혔다. [37689 : 2]

어린이와 가난한 이들의 건강을 위해 지칠 줄 모르게 질병과의 싸움에 몸을 내던진 위대한 사람은 마침내 평화와 안식의 길로 떠나 전설이 됐다. [37801 : 1]

고 이종욱 세계보건기구(WHO) 사무총장의 유해가 28일 오전 파리발 에어프랑스 AF264편으로 인천국제공항에 도착했다. [38032 : 1]

4. 결 론

온라인으로 보도되는 뉴스의 수가 급증하면서 다양한 뉴스 정보원으로부터 시차를 두고 생산되는 동일한 사건에 관련된 기사들로부터 종합적인 요약 정보를 얻고자 하는 이용자의 요구가 커지고 있다.

이 연구에서는 실시간으로 유입되는 뉴스 기사를 대상으로 같은 사건을 보도하는 기사들을 찾아내는 사건탐지 기법을 제안하고, 동일한 사건 관련 기사들로부터 하나의 요약문을 생성하였다. 이 연구에서 제안한 사건탐지 기법은 먼저 SVM 분류기를 이용하여 뉴스 기사를 사건 주제범주로 우선 분류한 후 주제범주 내에서 특정한 사건 관련 기사들의 클러스터를 생성하였다.

실험 결과 주제범주 분류 후 사건을 탐지한 경우 단순 클러스터링 기법만을 사용하였을 경우에 비해 오보율을 제외한 모든 척도에서 성능이 향상되었다. 또한 시간벌점함수와 클러스터 분할 기법을 싱글패스 클러스터링 알고리즘에 적용하였을 경우 F-척도와 탐지비용 측면에서 모두 성능이 향상된 것으로 나타났다.

이 연구에서 제안한 기법은 클러스터링 알고리즘만을 사용하는 사건탐지 기법에 비해 정확률, 재현율, F-척도에서 모두 우수한 성능을 보였으며, 사건탐지 기법의 주요 성능 척도인 누락률, 오보율, 탐지비용에서도 더 좋은 성능을 보였다. 사건 관련 뉴스 기사에서 시간, 장소, 인명 등 고유명의 역할이 매우 중요하지만 한국어 고유명 사전 등 적절한 자연언어 처리 도구를 찾을 수 없어서 이 연구에서는 고유명사에 가중치를 부여하는 방법을 대신 사용하였다. 그러나 통계적인 정보만을 이용한 자연언어 처리에서는 큰 성능 향상을 기대하기가 어렵다. 앞으로 한국어 대상의 자연언어 처리 기법의 발전에 따라 더 좋은 성능의 사건탐지 시스템 개발이 가능할 것으로 기대된다.

참 고 문 헌

정영미. 2005. 『정보검색연구』. 서울: 구미무역(주)출판부.

한국언론연구원. 1991. 『전국언론사 기사자료 표준분류표』. 서울: 한국언론연구원.

Allan, J., ed. 2002. *Topic Detection and Tracking Event based Information Organization.* Boston: Kluwer Academic Publishers.

Chang, C. and C. Lin. 2001. "LIBSVM: a library for support vector machines." Software available at 〈http://www.csie.ntu.edu.tw/~cjlin/libsvm〉.

Dharanipragada, D. M., J. S. Franz, McCarley, T. Ward, and W.-J. Zhu. 2002. "Segmentation and detection at ibm: hybrid statistical models and two-tiered clustering." In Allan, J eds. Topic Detection and Tracking Event based Information Organization. Kluwer Academic Publishers. 135-148.

Leek, T., R. Schwartz, and S. Sista. 2002. "Probabilistic approaches to topic detection and Tracking." In Allan, J eds. Topic Detection and Tracking Event based Information Organization. Kluwer Academic Publishers. 67-83.

McKeown, K. R., R. Barzilay, D. Evans, V. Hatzivassiloglou, J. L. Klavans, A. Nenkova, C. Sable, B. Schiffman, and S. Sigelman. 2002. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. *In Proceedings of Human Language Technology Conference 2002*(HLT 2002). San Diego, CA, USA.

Papka, R. and J. Allan. 1998. *On-line new event detection using single pass clustering*, Technical Report UM-CS-1998-021.

Radev, D. R., H. Jing, and M. Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. *NAACL/ANLP Workshop on Automatic Summarization.* 21-30.

Radev, D. R., H. Jing, M. Stys, and D. Tam, 2004. "Centroid-based summarization of multiple documents." *Information Processing & Management*, 40(3): 919-938.

Walker, S., S. E. Robertson, M. Boughanem, G. J. F. Jones, and K. Sparck Jones, 1998. Okapi at TREC-6 automatic ad hoc, VLC, routing, filtering and QSDR. *In Proceedings of the Sixth Text Retrieval Conference(TREC-6).* Available: 〈http://trec.nist.gov/pubs.html〉.

Yang, Y., T. Adult, T. Pierce, and C. W. Lattimer. 2000. "Improving text categorization methods for event tracking."

In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 65-72, GR. ACM Press, New York, US.

Yang, Y., J. G. Carbonell, and R. D. Brown. 1999. "Learning Approaches for Detecting and Tracking News Events." *IEEE Intelligent Systems*, July-August: 32-43.