자동 추출된 시간정보를 이용한 사건 클러스터링

김평[O] 남덕윤 최기석 맹성현*
한국과학기술정보연구원
*한국정보통신대학교
{pyung[O], dynam, choi}@kisti.re.kr, myaeng@icu.ac.kr

Event Clustering Using Automatically Extracted Temporal Information

Pyung Kim[O], Dukyun Nam, KiSeok Choi, SungHyun Myaeng*
Korea Institute of Science and Technology Information
*Information & Communications University

요 약

신문기사를 대상으로 사건 단위로 문서를 클러스터링 하기 위해서, 기존의 연구에서는 기사의 발행일 또는 기사의 내용만 사용하여 하나의 사건을 다른 사건과 구분하는 방법을 사용해 오고 있다. 하지만 사건의 전개가 시간 차이를 두고 진행되는 경우 또는 비슷한 시간대에 같은 범주에 속하는 사건이 발생하는 경우 기사의 발행일만 사용하여 사건 관련 기사를 구분하는 것은 한계가 있다. 본 연구에서는 한국어 신문기사를 대상으로 신문기사에 나타난 시간정보를 자동 추출하고, 이를 기사의 발행일을 기준으로 정규화 한 후 사용하여 사건 단위로 기사를 클러스터링 하는 방법을 개발하였다. 즉 한국어 신문 기사를 대상으로 기사에 나타난 시간 표현을 자동으로 추출한 후, 사건과의 유사도 비교에 사용함으로써 사건 단위 클러스터링의 정확도를 높이기 위한 방법을 제안한다.

1. 서 론

시간정보를 사용해서 사건 관련 기사를 분류하는 연구는, 외국의 경우 뉴스기사나 방송자료를 대상으로 사건을 지정하고 해당 사건에 대한 관련기사를 추적하거나 새로운 사건에 대한 기사를 자동탐지 하는 연구[1,2,3]로 진행되고 있다. 이 분야에서 사건은 " 특정 시간과 장소에 발생한 어떤 일" 로 정의된다[1]. 즉 사건이 발생한 시간과 장소에 따라 서로 동일한 사건인지 또는 다른 사건인지를 구분할 수 있다. 따라서 사건과 관련된 정확한 시간정보 추출은 사건 관련 여부를 결정하는 중요한 역할을 하는 것은 물론 기사를 시간의 발생 순으로 정렬하여 시간별 사건추이를 이해할 수 있게 해 준다. 예를 들면 " 발리 폭탄테러" 와 " 필리핀 연쇄 테러" 는 " 테러" 라는 하나의 범주에 속하지만 서로 다른 사건으로 구분할 수 있다. 사건을 기술하기 위해 사용되는 사건과 관련된 시간과 장소는 하나의 사건을 다른 사건과 구분하는데 중요한 역할을 한다. 본 연구에서는 한국어 신문기사를 대상으로 기사에 언급된 시간정보를 추출하고 정규화한 후, 시간정보와 사건과의 관련도에 따라 사건 단위의 클러스터링의 유사도 계산에 가중치를 부여함으로써 클러스터링의 정확도를 개선하는 것을 목표로 하였다.

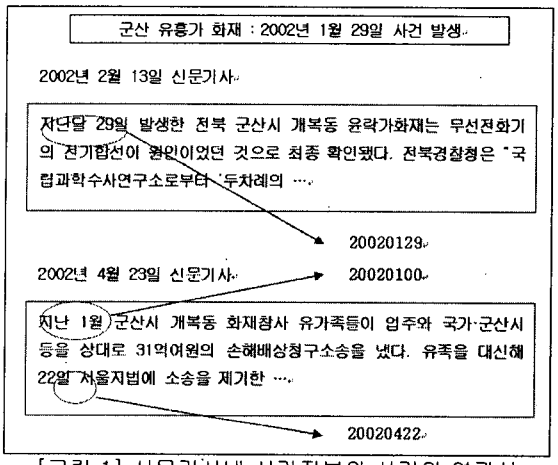
사건별로 기사들을 클러스터링하는 연구로는 그룹 평균 클러스터링 알고리즘을 사용해서 계층적 클러스터를 생성하거나 또는 단일 패스 점진적 클러스터링 알고리즘을 사용해서 비계층적 클러스터를 생성하는 과정에서 기사의 발행일 간의 차이를 유사도에 반영하는 연구[1], 점진적 K-means 알고리즘을 사용해서 기사들을 클러스터링 하는 연구등 다양한 연구들이 있다. 기사의 발행일간의 차이를 유사도에 반영하는 연구[1]는 일정한 시간 윈도우에 속한 기사들간의 발행일 차이 정도에 따라 유사도를 낮추는 함수(decay function)를 적용함으로써 발행일간의 차이가 나면 날수록 같은 사건으로 판정되는 것을 방지하였다.

2. 시간정보과 사건과의 연관성

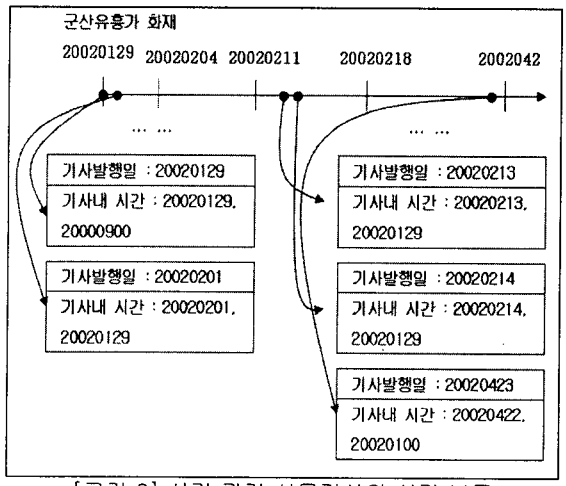
일반적인 경우 사건 발생일을 기준으로 1~2주 정도 기간에 사건 관련 기사가 모두 나타나게 된다. 하지만 일부 사건의 경우 사건의 진행에 따라 시간 차이를 두고 사건 관련 기사가 나타나는 경우도 많다. 이 경우 단순하게 기사의 발행일만 사용하여 사건 관련 여부를 판별하기가 어렵다.

[그림 1]은 2002년 1월 29일 발생한 ' 군산 유흥가 화재' 사건과 관련된 보도 기사 중 일부 기사를 보여주고 있다. 2002년 2월 13일 신문기사에서는 ' 지난달 29일 …' 이라는 표현을 통해 다른 사건과 이 사건을 구별할 수 있는 단서를 제공하고 있다. ' 지난달 29일' 이라는 표현은 기사의 발행일이 2002년 2월이므로 여기서 ' 지난달' 은 2002년 1월을 가리키고 있음을 알 수 있다. 따라서 ' 지난달 29일' 이라는 표현은 2002년 1월 29일이라는 것을 알 수 있으며 이를 ' YYYYMMDD' (연월일)로 표현하면 ' 20020129' 로 정규화 할 수 있고, 이는 ' 군

산 유흥가 화재' 의 사건 발생일과 일치하므로 같은 사건을 보도하고 있음을 알 수 있다. 사건 발생일을 기준으로 약 3달 가까이 시간적 차이를 두고 보도된 2002년 4월 23일 신문기사에서는 ' 지난 1월 …' 이라는 표현을 통해 이전 사건과의 관련성을 나타내고 있다. 기사의 발행일이 2002년 4월 23일이므로 ' 지난 1월' 은 2002년 1월을 나타내고 있음을 알 수 있고, 따라서 2002년 1월 29일 발생한 ' 군산 유흥가 화재' 와 연관성을 표현하고 있음을 알 수 있다.



[그림 1] 신문기사내 시간정보와 사건의 연관성



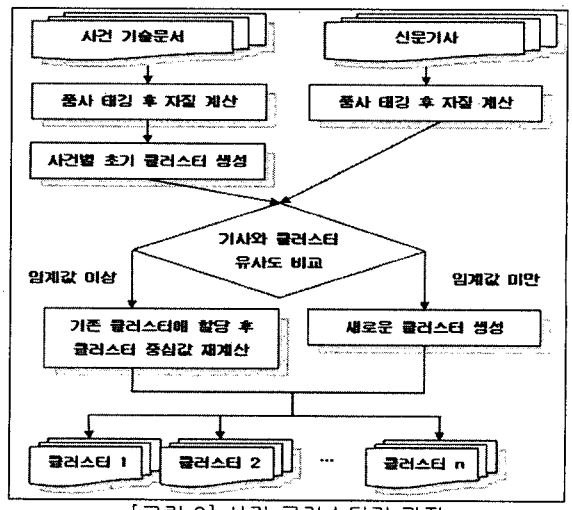
[그림 2] 사건 관련 신문기사의 시간 분포

[그림 2]는 ' 군산유흥가 화재' 사건의 관련기사의 시간별 분포에 따라 기사의 발행일과 기사에서 자동 추출된 시간을 보여주고 있다. 그림에서와 같이 대부분의 경우 기사에서 추출된 시간들이 사건 발생일과 일치하거나 사건 발생일을 포함하고 있음을 알 수 있다.

3. 사건 클러스터링

3.1 시스템 구성

사건별 클러스터에 사건 관련 기사를 할당하기 위한 처리 과정은 [그림 3]과 같다.



[그림 3] 사건 클러스터링 과정

일련의 신문기사가 시스템에 기사의 발행일 순으로 입력되면, 단일경로 클러스터링 알고리즘을 사용해서 기사의 내용 자질을 추출한다. 추출된 내용 자질을 사용하여 클러스터와 기사간의 내용 유사도를 계산한다. 하나의 기사는 사건 클러스터와 내용 유사도를 계산한 후 미리 지정된 임계값 이상인 모든 클러스터에 중복 할당된다. 하나의 클러스터에서는 속하는 기사에 한하여 기사내 절대시간과 상대 시간표현을 추출하여 정규화를 수행한 후 사건과의 연관도에 따라 가중치를 달리 부여하여 사건 기술문서와 최종 유사도를 비교하게 된다. 이런 과정을 통해 하나의 기사는 최종 유사도가 가장 높은 하나의 사건 클러스터에 할당되어 관련기사로 판정된다.

3.2 시간정보 추출

신문기사에서 시간정보를 추출하는 방법은 품사패턴과 어휘사전을 이용하는 방법[4]을 사용하였다. 유한오토마타와 어휘사전, 그리고 정규화 규칙을 적용해서 기사의 발행일을 기준으로 신문 기사내의 시간표현을 정규화하였다. 시간표현은 시각정보와 기간정보로 구분되어 정규화된다.

3.3 시간정보를 이용한 유사도 계산

기사와 사건의 최종 유사도는 기사의 내용 유사도에 기사의 발행일, 기사내 자동 추출 시간과 사건 발생일간의 유사도를 반영하여 계산된다. 하나의 기사는 최종 유사도 계산에 따라 하나의 사건 클러스터에 할당된다. 최종 유사도는 [수식 1]을 사용하여 계산되며 내용 유사도와 시간 유사도에 따른 상대적 가중치로 $\alpha = 0.6$와 $\beta = 0.4$ 값이 사건 문서에 대한 실험 문서 분석에 의해 결정되었다.

기사와 사건의 시간 근접도는 기사와 관련된 시간들에 부여된 우선 순위와 사건 발생일과의 시간적 근접도를 이용하여 계산되며 그 수식은 [수식 2]와 같다.

$$TotalSim(T_i, d_j) = \alpha \times SimContent(T_i, d_j) + \beta \times SimTime(T_i, d_j)$$

$TotalSim(T_i, d_j)$ = 사건 T_i 와 문서 d_j 간의 최종유사도

α = 내용 유사도에 대한 가중치

$SimContent(T_i, d_j)$ = 사건 T_i 와 문서 d_j 간의 내용 유사도

$SimTime(T_i, d_j)$ = 사건 T_i 와 문서 d_j 간의 시간 근접성

β = 시간 근접도에 대한 가중치

[수식 1] 기사와 사건의 최종 유사도

$$SimTime(T_i, d_j) = \Pr iority(d_j, t_k) \times AE(T_i, t_k)$$

$SimTime(T_i, d_j)$ = 사건 T_i 와 문서 d_j 간의 시간 유사도

$\Pr iority(d_j, t_k)$ = 문서 d_j 에서 시간 t_k 의 우선 순위에 따른 중요도

$AE(T_i, t_k)$ = 사건 T_i 와 시간 t_k 의 시간 근접성에 따른 가중치

[수식 2] 기사와 사건의 시간 근접도

사건과 문서간의 시간 근접성은 시간정보의 우선 순위와 사건 발생일과의 관련도에 따라 정해지며, 이 때 사용된 상수 값들은 사건 문서의 샘플 문서 분석을 통해 결정되었다. 시간정보에 대한 우선 순위 부여와 사건 발생일과의 관련도는 다음과 같은 기준에 의해 결정된다.

- 기사의 발행일이 사건 발생일 이전인 경우 최종 유사도 계산에서 제외
- 기사와 관련된 시간의 중요도에 따른 우선 순위에 따라 조정
- 우선 순위가 낮을수록 단계별로 가중치를 낮춘다.
  - 1 순위 : 1.0, 2 순위 : 0.9, …
- 기사와 관련된 시간과 사건 발생일간의 근접도에 따른 우선순위
  - 사건 발생일과 일치, 하루 차이 또는 사건 발생일을 포함하는 경우 : 1.0
  - 사건 발생일 1주일 안에 해당되는 경우 : 0.9
  - 사건 발생일 2주일 안에 해당되는 경우 : 0.7
  - 그 외의 경우 : 0.5

4. 실험

시간 추출 시스템의 평가는 2002년도 조선일보 신문기사 12만 건에서 미리 선정된 25개의 사건, 1320건의 관련기사를 대상으로 실험하였다. 각각의 사건은 사건에 대한 발생일과 간단한 설명문서를 포함하고 있으며, 이 설명문서를 사용하여 사건별 초기 클러스터를 생성하게 된다.

[표 1]은 사용된 25개의 사건리스트와 사건별 정답기사수를 보여주고 있다. [표 1]에 제시된 25개의 사건 클러스터에 사건과 관련된 기사를 할당하는 과정에서 기사의 내용과 발행일을 사용하는 경우(A), 기사의 내용, 발행일, 기사에서 자동 추출된 시간정보를 동일하게 사용하는 경우(B), 기사의 내용, 발행일, 기사에서 자동 추출된 시간을 가중치 부여 방식에 따라 계산한 경우[C]로 구분하여 실험하였다.

[표 1] 사건 리스트

번호	사건명	정답기사수
1	군산시 유흥가 화재	22
2	미국 악의축 국가로 북한 선언	141
3	김동성-오노 동계올림픽 판정 시위	81
4	서울 상봉동 은행에 총기 강도 사건	17
…	…	…
25	손기정옹 타계	20
계		1320

[표 2] 클러스터링 결과 비교

유형	정확율	재현율	F-measure
(A)	70 %	60 %	65 %
(B)	75 % (+7 %)	72 % (+20 %)	73 % (+12 %)
(C)	81 % (+16 %)	70 % (+17 %)	75 % (+ 15 %)

[표 2]는 각각의 조건별 클러스터링 재현율과 정확도를 보여주고 있다. 기사의 발행일만 사용한 경우에 비해 기사에서 자동 추출된 시간을 사용한 경우가 F-measure 에서 12% 정도의 성능 향상을 보였고, 시간 정보와 사건 발생일과의 관계에 따라 차등 유사도를 적용한 경우 약 15 % 정도의 성능 향상을 가져왔다.

5. 결론

본 논문에서는 사건 단위의 클러스터링 성능을 향상 시키기 위해 기사에서 자동 추출된 시간정보를 사용하는 방법을 제시하였다. 그리고 기사의 발행일을 사용하는 경우와 기사의 발행일과 기사에서 추출된 시간을 사용하는 경우를 비교하여 실험함으로써 사건 클러스터링 과정에서 자동 추출된 시간 정보의 유용성을 증명하였다. 또한 자동 추출된 시간과 사건과의 관련도, 추출된 시간과 사건의 발생일간의 관련 유형에 따라 우선 순위를 정하고 유사도를 차등 적용하여 토픽 추적 과정의 정확도를 향상시켰다. 향후 연구로는 기사에서 추출된 시간을 사용하는 경우 사건과 관련되지 않은 시간정보가 추출되어 발생하는 오류를 줄이기 위한 방법으로 추출된 시간이 사건과 관련이 있는지를 판단할 수 있는 방법이 요구된다.

참고문헌

[1] Y. Yang외 2, " A Study on Retrospective and On-Line Event Detection." , In Proceedings of ACM SIGIR Conference on Research and development in information retrieval, pp 28-36, 1998.

[2] J. Allan외 2, " On-line new event detection and tracking." , In Proceedings of ACM SIGIR conference on Research and development in information retrieval, pp 37 - 45, 1998.

[3] J. Allan. " Introduction to Topic Detection and Tracking." , TOPIC DETECTION AND TRACKING, Kluwer Academic Publishers, pp 1-16. 2002.

[4] P. Kim, S. H. Myaeng "Usefulness of Temporal Information Automatically Extracted from News Articles for Topic Tracking.", ACM TALIP, Vol. 3, No. 4, pp 227-242, 2004