사건트래킹과 정보필터링 기법의 사건검색 성능 비교연구

A Comparative Study on the Event-Retrieval Performances of Event Tracking and Information Filtering

정영미(Young Mee Chung)*, 장지은(Jieun Chang)**

초 록

이 연구의 목적은 사건을 연구대상으로 하는 사건트래킹 기법이 과연 최신 사건 정보를 검색함에 있어 기존의 정보필터링 기법보다 성능이 우수한가를 살펴보는 데 있다. 따라서 이 연구에서는 특정 사건에 관한 최신 기사를 보다 효과적으로 검색하여 제공하는 기법을 찾아내기 위하여 kNN(k-Nearest Neighbors) 분류기를 응용한 사건트래킹 기법과 질의기반 정보필터링 기법을 사용하여 사건검색 실험을 수행한 후 두 기법의 검색 성능을 비교하였다. 사건트래킹 실험은 초기의 고정 학습문서 집합을 사용한 사건트래킹과 트래킹 과정에서 변화하는 동적 학습문서 집합을 사용한 사건트래킹의 두 가지 방법으로 수행되었다. 정보필터링 실험도 초기질의를 사용한 정보필터링과 필터링 과정에서 계속 수정되는 질의를 사용한 정보필터링의 두 가지 방법으로 수행되었다. 실험 결과 사건트래킹 기법에서는 고정 학습문서 집합을 사용한 경우가 동적 학습문서 집합을 사용한 경우보다 더 우수한 성능을 보였으며, 정보필터링 기법에서는 초기질의를 사용한 경우가 수정질의를 사용한 경우보다 더 좋은 성능을 보였다. 또한 고정 학습문서 집합을 사용한 사건트래킹과 초기질의를 사용한 정보필터링을 비교한 결과 정보필터링 기법이 사건트래킹 기법에 비해 더 좋은 사건검색 성능을 보이는 것으로 나타났다.

ABSTRACT

The purpose of this study is to ascertain whether event tracking is more effective in event retrieval than information filtering. This study examined the two techniques for event retrieval to suggest the more effective one. The event-retrieval performances of the event tracking technique based on a kNN classifier and the query-based information filtering technique were compared. Two event tracking experiments, one with the static training set and the other with the dynamic training set, were carried out. Two information filtering experiments, one with initial queries and the other with refined queries, were also carried out to evaluate the event-retrieval effectiveness. We found that the event tracking technique with the static training set performed better than one with the dynamic training set. It was also found that

* 연세대학교 문헌정보학과 교수(ymchung@yonsei.ac.kr)

** 연세대학교 문헌정보학과 대학원(jejang97@hotmail.com)

■ 논문 접수일 : 2003. 8. 20

■ 게재 확정일 : 2003. 9. 3

the information filtering technique using intial queries performed better than one using the refined queries. In conclusion, the comparison of the best cases of event tracking and information filtering revealed that the information filtering technique outperformed the event tracking technique in event retrieval.

키워드: 사건검색, 사건트래킹, 정보필터링, 검색성능, event retrieval, event tracking, information filtering, retrieval effectiveness

1 서 론

인터넷 환경에서 급격하게 증가하는 정보에 효과적으로 접근하기 위한 여러 검색 기법들이 연구되고 있으며, 이러한 연구들은 대부분 주제(topic) 검색에 그 초점을 맞추고 있다. 그러나 21세기 지구촌 시대가 도래함에 따라 세계 각처에서 발생하는 사건(event)과 이슈(issue)의 수는 급격히 증가하게 되었고 이에 대한 관심 또한 커지고 있다. 이제 한 개인이 시시각각 발생하는 뉴스를 모두 인지하는 것은 불가능하게 되었으며, 이에 따라 이용자가 자신의 관심사에 따라 새로운 사건들을 효과적으로 인지하고 검색할 수 있는 기법에 대한 연구가 필요하게 되었다.

1996년 미국정부에 의해 시작된 TDT (Topic Detection and Tracking) 연구 프로젝트는 현실에서 발생하는 일들을 기존의 검색 대상이 되어온 '주제(topic)'와 구별하여 '사건(event)'이라 지칭하고 이에 연구의 초점을 맞추고 있다. 즉, 주제와 구별되는 사건의 특성을 연구하며 새로운 사건 검색 기법들을 제안하였는데 이 가운데 하나가 사건트래킹(event tracking) 기법이다.

사건트래킹 시스템은 이용자가 관심이 있는 사건에 관한 문서를 시스템에 제시하면 이 문서를 바탕으로 새로 유입되는 정보 가운데 이와 유사한 사건 정보를 검색하여 이용자에게 제공한다. 사건트래킹은 이용자가 제시한 정보에 기초하여 이와 유사한 최신정보를 검색하여 지속적으로 이용자에게 제공한다는 점에서 정보필터링(information filtering)과 그 기능이 유사하다. 따라서 기존의 정보필터링 기법을 새롭게 발생하는 사건들의 검색에 적용하면 사건트래킹과 동일한 기능을 수행할 수 있다.

이 연구의 목적은 과연 새로 제안된 사건트래킹 기법들이 기존의 정보필터링 기법보다 사건검색에 있어 더욱 효과적인가를 확인하는 데 있다. 따라서 이 논문에서는 TDT 연구에서 제안한 kNN(k-Nearest Neighbor Classifier) 분류기 기반 사건트래킹 기법을 사용하여 사건검색을 실시하였고, 또한 동일한 실험집단 상에서 질의 기반 정보필터링 기법을 사용한 사건검색을 실시한 다음 두 기법의 검색 성능을 비교하였다. 이때 사건트래킹과 정보필터링 기법은 각각 트래킹과 필터링 과정에서 처리된 문서를 활용하는지의 여부에 따라 서로 다른 두 가지 방법으로 실험하였다. 즉, 사건트래킹 기법에 대해서는 초기 학습문서 집합을 고정적으로 사용하는 실험과 트래킹 과정에서 긍정으로 판정된 문서를 긍정 학습문서로 포함하는 동적 학습문서 집합을 사용하는 실험을 수행하였다. 정보필터링 기법에 대해서는 초기 질의를 사용하는 실험과 필터링 과정에서 긍정으로 판정된 문서의 정보에 의해 수정된 질의를 사용하는 실험을 각각 수행하였다.

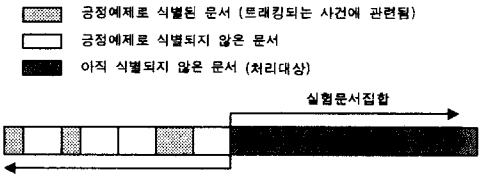
2 사건트래킹과 정보필터링의 개념

2.1 사건트래킹(event tracking)

엄청난 양의 뉴스를 통제하기 위하여 미국 정부의 DARPA(Defense Advanced Research Projects Agency)에 의해 1996년 시작된 TDT 연구에서는 기존의 검색 대상이 되어 온 주제와 구별하여 사건의 개념을 처음으로 도입하였다. TDT 연구에서는 사건이란 현실 세계에서 발생하는 구체적이고 실제적인 일들을 지칭하며, 하나의 사건은 그것과 직접적으로 관련되는 모든 활동을 포함한다고 정의한다(Allan 2002). 예를 들어 어떤 폭발 사건이 발생한 경우, 그 폭발 자체는 이후 일련의 이야기들을 유발하는 최초의 사건 예제가 된다. 또한 그 폭발에 관한 내용을 비롯하여 구출 작업, 범인 색출, 처벌 및 재판 등의 이후 이야기들도 해당 사건에 포함된다.

사건은 전통적인 개념의 주제에 비하여 그에 속하는 정보의 수가 많지 않다. 어느 정도 시간이 경과하면 그에 대한 정보는 더 이상 생산되지 않으며 이와 같은 특징은 사건을 주제와 구별시켜 주는 중요한 요인이 된다. 예를 들어 주제적인 성격을 가지는 지하철 사고는 시간의 흐름에 관계없이 지하철에서 발생하는 모든 종류의 사고를 지칭한다. 반면 대구 지하철 화재 참사 사고는 특정 시간에 특정 장소에서 특정 인물에 의해 발생한 사고를 가리키며 시간의 경과에 따라 이와 관련된 정보는 더 이상 생산되지 않는다.

사건트래킹은 TDT의 대표적인 세부 연구과제 중 하나로, 시스템 상에 순차적으로 존재하는 문서들에 대하여 사전에 정의된 사건 명칭을 자동으로 부여해 주는 작업을 말한다(Yang et al. 2000). 즉, 사건트래킹은 시스템이 사전에 인식한 사건과 새로 들어오는 이야기들을 연결시켜 주는 작업이라 할 수 있다(Allan, Lavrenko, and Papka 1998). 〈그림 1〉은 사건트래킹의 개념을 보여준다(National Institute of Standards and Technology 2000).



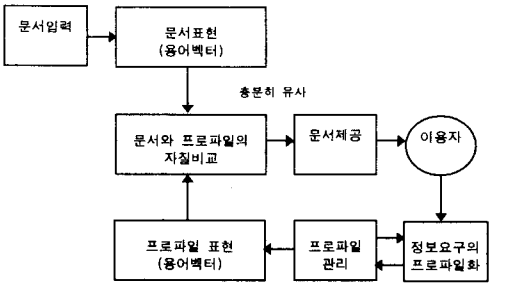
〈그림 1〉 사건트래킹의 개념

사건트래킹 실험은 보통 실험집단을 학습문서 집합과 실험문서 집합의 두 개로 분리한 뒤 일련의 학습과정을 거쳐 수행된다. 이때 분리 기준은 대상 사건이 발생하고 그에 관련된 정보가 N 건 출현하는 시점으로 삼는데, TDT 연구에서는 주로 2, 4, 8, 16을 N 값으로 설정하여 학습문서 집합과 실험문서 집합을 분리한다(Papka 1999). 트래킹의 대상이 되는 각 사건은 N 값에 따라 서로 다른 학습문서 집합과 실험문서 집합을 가지게 된다. 학

습문서 집합에 속하는 문서는 해당 사건의 내용을 담고 있는지의 여부가 표시되며, 이는 시스템이 그 사건을 트래킹 할 수 있는 유일한 정보가 된다. 사건트래킹은 이 정보를 바탕으로 유입되는 실험문서가 해당 사건을 그 내용으로 하는지의 여부를 판별하는 것이다(Allan 2002). 이때 각 사건에 대한 트래킹은 서로 독립적으로 이루어지며, 따라서 하나의 사건에 대한 문서의 판별은 다른 사건에 대한 그 문서의 판별에 영향을 끼치지 않는다(Allan, Lavrenko, and Papka 1998). 이 연구에서는 예제기반 범주화 기법인 kNN 분류기를 사용하여 사건트래킹을 수행하였다.

2.2 정보필터링(information filtering)

정보필터링은 끊임없이 생산되는 정보 중 이용자에게 필요한 최신정보만을 제공하기 위한 정보검색 기법으로서 대부분의 정보필터링 시스템은 이용자로 하여금 프로파일 등을 통하여 자신의 관심 주제를 제시하도록 한다. 이용자가 제시한 프로파일과 새로 유입되는 문서는 흔히 용어 벡터 형태로 변환된 다음 두 벡터간의 유사도가 측정되며, 프로파일과의 유사도가 기준치 이상인 문서들이 이용자에게 제공된다. 개략적인 정보필터링 과정은 〈그림 2〉와 같다.



〈그림 2〉 정보필터링 과정 개요

이 연구에서는 여러 필터링의 방식 가운데 적응적 필터링(adaptive filtering)을 선택하여 실험에 사용하였다. 즉, 이용자가 관심을 가지는 사건을 내용으로 하는 기사들을 시스템에 제시하면서 필터링이 시작되며 그 이외의 어떠한 정보도 사용되지 않는다. 그러나 적응적 필터링에서는 일반적 필터링 과정에서 이미 처리된 문서 정보를 사용해 프로파일을 갱신하는 것과 마찬가지로 처리된 문서 중 긍정문서의 정보를 활용해 질의를 계속 수정할 수 있다.

3 사건트래킹과 정보필터링 기법의 실험

3.1 실험설계

현실 세계에서 발생하는 구체적이고도 실제적인 사건을 연구하기 위해서는 주로 신문이나 방송의 뉴스를 문자화하여 실험 집단으로 사용한다. 본 연구에서는 중앙일보에서 웹을 통하여 제공하는 기사들을

〈표 1〉 사건명과 실험집단

번 호	사 건 명	학습문서 수 (긍정문서 수)	실험문서 수 (긍정문서 수)
1	주한미군 재배치 및 감축논의	3,087(4)	6,937(202)
2	김대중 정권 대북 지원설	806(4)	9,218(649)
3	대구 지하철 화재 참사 사고	5,504(4)	4,520(436)
4	SK 글로벌 분식회계 비리 사건	5,411(4)	4,613(88)
5	개그우먼 이경실 폭행 사건	4,576(4)	5,448(8)
6	고건 국무총리 인증	972(4)	9,052(78)
7	노무현 정권 검찰 공무원 개혁 단행	4,616(4)	5,408(64)
8	임동원 국정원장 대북 특사파견	1,759(4)	8,265(59)
9	남북 장관급 회담	1,084(4)	8,940(38)
10	인터넷 대란 사건	1,928(4)	8,096(43)
총 문서수		10,024	

수집하여 자체적으로 실험집단을 구축하였다. 중앙일보 2003년 1월 15일~2003년 3월 15일까지의 정치면과 사회면 기사 총 10,024 건이 실험집단으로 사용되었으며, 실험집단에 나타난 사건들 중 10개의 사건을 임의로 선정하여 트래킹과 필터링 실험을 수행하였다.

이 연구에서 채택한 사건트래킹과 정보필터링 기법은 모두 학습에 기반한 기법이므로 전체 실험집단을 학습문서 집합과 실험문서 집합으로 구분하였다. 학습문서들은 학습을 위해 사용되며, 실험문서들은 트래킹과 필터링 실험을 위해 새로 입력되는 문서의 역할을 하게 된다. 고정된 학습문서 집합을 사용하는 실험에서는 해당 사건과 관련된 기사가 4번 출현하는 시점을 기준으로 하여 그 이전의 문서를 학습문서 집합으로 삼고 이후의 문서를 실험문서 집합에 포함시켰다. 이때 해당 사건과 관련된 기사는 긍정문서로 간주한다. 따라서 각각의 사건에 대해 모두 고유한 학습문서 집합과 실험문서 집합 상에서 트래킹과 필터링이 실시되었다. 선정된 사건과 그에 따른 학습문서와 실험문서의 수, 그리고 각 집합에 포함된 긍정문서의 수는 〈표 1〉과 같다.

사건트래킹과 정보필터링 기법의 검색실험을 위해 벡터공간 검색 모형을 선택하였다. 벡터공간 검색은 문서와 질의를 각각 용어벡터 형식으로 표현하고 두 벡터 사이의 유사도를 산출한 다음, 유사도 값이 기준치 이상인 문서를 유사도의 순서에 따라 제공하는 검색 기법이다 (Salton 1983). 용어벡터를 구성하는 용어들의 가중치로는 코사인 정규화된 $tf \cdot idf$ 공식을 사용하였다. 즉, 학습집단 내의 문서 수가 N 이고 용어 t 가 출현한 학습문서의 수가 n_t, 문서 내에서의 단어빈도가 $tf(t, d)$ 라면 단어 t 가 문서 d에서 가지는 가중치 $w(t, d)$는 다음과 같이 산출된

다(Yang et al. 2000).

$$w(t,d)=\frac{(1+\log_2 tf(t,d))\times\log_2(N/n_t)}{\sqrt{\sum_{t\in d}[(1+\log_2 tf(t,d))\times\log_2(N/n_t)]^2}}$$

이 연구에서는 TDT 연구의 성능 평가 방법을 적용하여 각 기법의 사건검색 성능을 평가하였다. 이를 위해 각 사건마다 〈표 2〉와 같은 분할표가 만들어졌으며, 이를 바탕으로 하여 누락률, 오보율, 재현률, 정확률, 트래킹비용의 5가지 척도를 이용하여 사건검색 성능을 측정하였다. 누락률은 검색되지 않은 적합문서의 비율을, 오보율은 전체 부적합문서 중 검색된 부적합문서의 비율을 나타내며, 트래킹 비용은 적합문서를 부적합문서로, 또 부적합문서를 적합문서로 잘못 판정함으로써 발생하는 비용을 의미한다. 이 세 가지 척도는 모두 값이 작을수록 좋은 검색 성능을 나타낸다.

〈표 2〉 시스템의 성능평가를 위한 분할표

	실제로 긍정	실제로 부정
시스템이 긍정으로 예상	*a*	*b*
시스템이 부정으로 예상	*c*	*d*

TDT 연구에서는 사건트래킹에서 문서가 잘못 처리되는 비율을 우선적인 기준으로 하여 성능을 평가하였는데 그 중에서도 트래킹비용을 최우선 척도로 삼았다. 〈표 2〉에 기반한 평가 척도 공식은 다음과 같다(Yang et al. 2000).

■ 누락률(miss) : $m=\frac{c}{a+c}\quad(a+c>0)$

■ 오보율(false alarm) : $f=\frac{b}{b+d}$ $(b+d>0)$

■ 재현율 (recall) : $r=\frac{a}{a+c}\quad(a+c>0)$

■ 정확률(precision): $p=\frac{a}{a+b}\quad(a+b>0)$

■ 트래킹 비용(tracking cost): $\alpha_1\frac{b}{n}+\alpha_2\frac{c}{n}$ $(n=a+b+c+d)$

본 연구에서도 TDT 연구에서와 같이 트래킹비용을 최우선 평가척도로 삼았으며, 트래킹비용의 파라미터 값도 $\alpha_1=0.1$, $\alpha_2=1$로 하였다. 이 값은 트래킹비용 산출시 적합문서가 부적합문서로 잘못 판정되는 비율이 부적합문서가 적합문서로 잘못 판정되는 비율보다 더 많이 반영됨을 의미한다.

이 연구에서는 위의 5가지 평가척도 이외에 재현율과 정확률을 결합한 단일가 척도인 F1 척도가 추가적으로 사용되었으며, 다음 공식에서와 같이 재현율과 정확률의 중요성을 동일하게 반영하여 F1 값을 산출한다.

$$F_1=\frac{2rp}{r+p}$$

3.2 사건트래킹 실험

3.2.1 고정 학습문서 집합을 사용한 사건 트래킹

kNN 분류기를 사용하는 사건트래킹 기법은 용어벡터로 표현된 학습문서와 입력문서를 서로 비교하여 하나의 입력문서에 대하여 가장 유사한 k 개의 이웃문서를 학습문서 집합으로부터 선정한 후, 선정된 이웃문서들의 사건 범주정보를 이용하여 트래킹을 실시한다. 이때 학습문서와 입력문서 간의 유사도는 코사인 유사계수를 통해 계산되며, 선정된 k 개의 이웃문서 중 긍정 학습문서들과의 유사도와 부정 학습문서들과의 유사도 차에 의해 입력문서에 대한 최종 결정값이 구해진다. 입력문서에 대한 트래킹은 이 결정값이 기준치를 넘을 때 이루어진다(Yang et al. 2000).

그러나 대개 하나의 사건이 극히 소수의 긍정예제만을 가지고 있기 때문에 이웃 문서에 긍정예제들이 충분히 포함되지 못하는 경우 유사도 합의 차이에 의해 결정값을 구하는 방식은 문제가 있다. 따라서 유사도 합의 차이 대신 유사도 평균값의 차이로 결정값을 구하는 방법이 긍정문서보다 부정문서들이 상대적으로 많아 그 값이 우세해 지는 것을 방지해 주며, 이 변형된 사건트래킹 기법이 원형의 kNN 사건트래킹에 비해 더 향상된 성능을 갖는 것으로 나타났다(Yang et al. 2000).

이 연구의 사건트래킹 실험에서는 유사도 평균값을 이용하여 트래킹 결정값을 구하는 방법을 사용하였다. 사건트래킹 시스템에서는 각 사건에 대한 트래킹이 독립적으로 이루어지며 학습문서 집합도 트래킹의 대상이 되는 사건에 해당하는가의 이진결정(긍정/부정) 정보만을 제공한다. 따라서 입력문서와의 자질 비교 결과 선정된 k개의 학습문서는 입력문서가 포함하는 사건을 다루고 있는 긍정문서와 그렇지 않은 부정문서로 나뉘게 된다.

학습문서 집합이 D이고 선정된 k 개의 이웃문서 중 긍정문서 집합이 P_k, 부정문서 집합이 Q_k 일 때, 입력문서 x의 결정값 r'은 다음과 같이 긍정문서 집합과 부정문서 집합간의 유사도 평균값의 차이를 나타낸다.

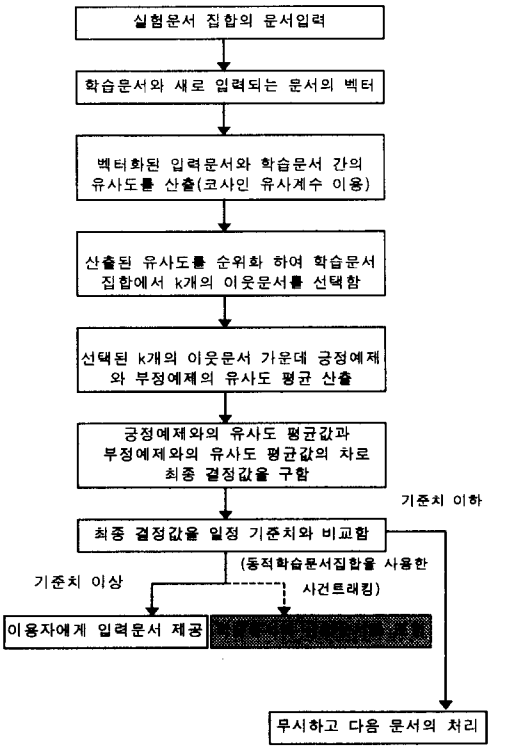
$$r'(x,k,D)=\frac{1}{|P_k|}\sum_{y\in P_k}\cos(x,y)-\frac{1}{|Q_k|}\sum_{z\in Q_k}\cos(x,z)$$

트래킹 실험에서 설정할 파라미터 값으로는 트래킹 여부를 결정하는 결정값 기준치와 학습문서 집합에서 선정되는 이웃문서의 수인 k가 있다. TDT 트래킹 연구에서는 각 사건마다 별도로 기준치를 최적화 하는 것이 어렵다고 보고 모든 사건에 대해 결정값 기준치를 동일하게 적용하고 있다(Allan 2002). Kurt(2001)는 사건트래킹에서 기준치를 결정하기 위해 10개의 서로 다른 기준치 값과 그에 따른 트래킹 비용을 산출한 다음 실험대상이 된 모든 사건의 트래킹비용 평균값을 구하고 그 가운데 가장 작은 트래킹비용을

갖는 값을 최적의 기준치로 선정하였다. 본 실험에서도 10개의 서로 다른 기준치를 설정하고 그에 따른 성능을 측정한 결과 최소의 트래킹비용 평균값을 갖는 0.0015를 트래킹 기준치로 선정하였다.

Yang 등(Yang et al. 2000)은 15,864건의 기사로 구성된 TDT-1의 실험집단 상에서 본 실험과 동일한 방법으로 사건트래킹을 수행할 때 최적화된 k 값으로 5를 사용한 바 있다. 본 실험에서는 이를 근거로 하여 k 값으로 각각 3, 5, 7, 10을 사용하여 예비 실험한 결과 최소의 트래킹비용 평균값을 산출한 10을 k 값으로 선정하였다.

고정 학습문서 집합을 사용한 사건트래킹의 실험 과정은 〈그림 3〉과 같다.



〈그림 3〉 사건트래킹 과정

3.2.2 동적 학습문서 집합을 사용한 사건트래킹

이 실험에서는 트래킹 과정에서 시스템에 의해 적합문서로 판정된 문서를 이전의 학습문서 집합에 긍정 학습문서로 추가하여 새로운 학습문서 집합을 구성하고 이를 사용한 사건트래킹을 수행하였다.

트래킹 과정에서 시스템이 긍정으로 판정한 문서를 학습문서 집합에 긍정문서로 포함시킨다면 시간의 경과에 따른 사건의 변화 내용을 트래킹 과정에 반영할 수 있다. 또한 초기의 긍정 학습문서에 나타나지 않은 사건의 세부적인 내용을 트래킹 과정에서 추가적으로 반영할 수 있다는 장점이 있다(Allan, Larvrenko, and Papka 1998).

그러나 잘못 검색되어 긍정 학습문서에 포함되는 부적합 문서들의 내용이 계속 트래킹 과정에 반영된다면 해당 사건과는 관련 없는 문서들이 검색될 가능성이 크다. 즉 시스템에 의해 내려진 판정을 기준으로 실험문서를 긍정 학습문서 집합에 포함시키기 때문에 초기 트래킹의 성능이 좋지 못한 경우에는 점차적으로 트래킹의 성능이 더 나빠질 수 있다. 이와 같은 경우 트래킹 전체의 성능에 부정적인 영향을 미치게 된다.

동적 학습문서 집합을 사용한 사건트래킹의 실험과정은 〈그림 3〉에서와 같이 고정 학습문서 집합을 사용한 사건트래킹과 전반적으로 동일하며 다만 회색으로 표시된 작업이 추가되어 있다.

3.3 질의기반 정보필터링 실험

3.3.1 초기질의를 사용한 정보필터링

이 실험에서는 필터링의 대상이 되는 사건과 관련된 긍정예제들이 시스템에 제시되면 이로부터 핵심어를 추출하여 질의를 생성한다. 핵심어를 선정하는 방법으로는 Carpineto(2001)가 KLD(Kullback-Leible Divergence) 정보이론을 응용하여 고안한 공식을 사용하였다. 이 공식에서 단어 t가 특정문서 r에 출현할 확률을 $p_r(t)$라고 하고 전체 문서집단에 출현할 확률을 $p_c(t)$라고 할 때 단어 t가 특정문서 r에서 갖는 중요도 $score(t)$는 다음과 같이 산출된다.

$$score(t)=[p_r(t)]\times[p_r(t)/p_c(t)]$$

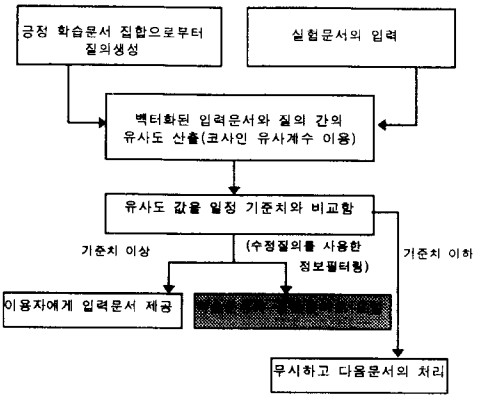
이 공식을 통해 산출된 중요도를 기준으로 각 단어를 순위화한 다음 상위에 오는 단어들로 각 사건에 대한 질의벡터를 생성한다. 질의벡터를 구성하는 최적의 질의어 수를 결정하기 위해 30, 50, 70, 100개씩 질의어를 추출하여 예비실험을 수행하고 그 결과 최적의 성능을 보인 100개를 질의어의 수로 선정하였다.

이 실험에서는 각 입력문서에 대하여 질의와의 유사도를 코사인 유사계수 공식을 통해 구하고, 유사도 값이 기준치를 넘는 입력문서는 해당 사건으로 필터링하였다. 이때 사건트래킹 실험에서와 같이 모든 사건에 대하여 동일한 기준치를 적용할 경우 필터링 시스템의 성능이 매우 낮아지는 것으로 나타났다. 정보필터링에서는 긍정문서의 수가 많은 사건의 경우에는 낮은 기준치가 적용되어야 하고 긍정문서의 수가 적은 경우에는 높은 기준치가 적용되는 것이 바람직하다. 따라서 이 실험에서는 각 사건마다 30개의 학습문서를 대상으로 한 예비실험에서 별도로 산출한 최적의 기준치를 적용하였다.

초기질의를 사용한 정보필터링 실험 과정은 〈그림 4〉에 나와 있다.

3.3.2 수정질의를 사용한 정보필터링

이 실험에서는 지역적 분석(local analysis)에 기반한 질의확장 기법을 응용하여 질의를 수정하고, 이 수정된 질의를 바탕으로 정보필터링을 수행한 다음 그 성능을 초기질의를 사용한 정보필터링과 비교하였다. 지역적 분석은 초기 검색 후 상위 순위의 문서들을 사용해 질의를 수정하는 방법으로서 검색된 상위 문서에 출현한 용어들을 자동적으로 질의에 추가한다. 이와 같은 방법의 질의확장은 초기 검색 결과를 이용하기 때문에 초기 검색 결과의 적합성에 따라 그 성능이 좌우된다는 특징을 갖는다(Xu and Croft 1996).



〈그림 4〉 정보필터링 실험 과정

질의의 수정은 하루 단위로 이루어졌는데, 필터링이 진행되면서 시스템에 의해 적합하다고 판정된 문서들을 긍정 학습문서 집합에 포함시키고 이로부터 핵심어를 재선정하였다. 이때 핵심어를 추출하는 공식은 초기질의를 생성할 때와 동일한 공식을 사용하였다. 수정질의를 사용한 정보필터링의 실험 과정은 〈그림 4〉에서와 같이 전반적인 과정은 초기질의를 사용한 정보필터링과 동일하며 다만 회색으로 표시된 작업이 추가되어 있다.

4 실험 결과 분석

4.1 고정 학습문서 사건트래킹과 동적 학습문서 사건트래킹의 성능 비교

〈표 3〉은 각각 고정 학습문서 집합과 동적 학습문서 집합을 사용한 사건트래킹 실험의 검색 성능을 보여준다. 누락률을 살펴보면 고정_사건트래킹이 0.4672, 동적_사건트래킹이 0.6705로서 동적_사건트래킹의 경우 적합문서가 누락되는 비율이 더 높다는 것을 확인할 수 있다. 오보율은 고정_사건트래킹이 0.0130, 동적_사건트래킹이 0.0611로서 동적_사건트래킹이 부적합문서를 적합문서로 잘못 처리한 경우가 더 많은 것으로 드러났다.

검색 효율의 기본 척도인 재현율과 정확률에 있어서는 고정_사건트래킹이 동적_사건트래킹에 비해 성능이 더 좋은 것으로 나타났다. 단일가 척도인 F_1 값도 고정_사건트래킹이 0.4056, 동적_사건트래킹이 0.1191로 고정_사건트래킹의 값이 훨씬 높게 나타나 있다.

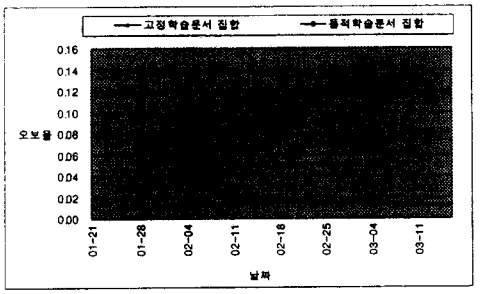
〈표 3〉 고정_학습문서_사건트래킹과 동적_학습문서_사건트래킹의 성능 비교

척 도	고정 학습문서 집합	동적 학습문서 집합
누 락 률	0.4672	0.6705
오 보 율	0.0130	0.0611
재 현 율	0.5328	0.3295
정 확 률	0.3977	0.1084
F1	0.4056	0.1191
트래킹비용	0.0141	0.0253

사건트래킹의 최우선 척도인 트래킹비용을 기준으로 살펴보면 고정_사건트래킹이 0.0141, 동적_사건트래킹이 0.0253으로서 트래킹 비용이 적은 고정_사건트래킹이 더 좋은 성능을 보였다.

Allan 등(Allan, Lavrenko, and Papak 1998)의 연구에 따르면 시스템의 오보율이 낮을 경우, 동적 학습문서 집합을 사용하는 사건트래킹 방법은 고정 학습문서 집합을 사용한 사건트래킹 방법보다 향상된 성능을 가질 수 있다. 그러나 오보율이 높은 경우에는 트래킹 과정에서 적합문서로 잘못 판정된 다수의 부적합문서가 학습문서 집합에 긍정문서로 포함됨에 따라 트래킹 시스템은 유입되는 실험문서에 대하여 잘못된 판정을 내리게 된다. 또한 시간의 경과에 따라 긍정 학습문서 집합에 포함되는 부적합문서가 증가하게 되므로 오보율은 지속적으로 증가한다. 동적 학습문서 집합을 사용한 사건트래킹 실험에서도 시간이 경과함에 따라 오보율이 점차 증가하였으며 이는 사건트래킹의 전반적인 성능에 좋지 못한 영향을 끼친 것으로 판단된다.

〈그림 5〉는 고정 학습문서 집합을 사용한 사건트래킹의 오보율은 시간의 경과에 따라 큰 차이를 보이지 않는 반면 동적 학습문서 집합을 사용한 사건트래킹의 오보율은 지속적으로 증가하는 것을 보여준다. 이와 같은 오보율의 지속적인 증가로 인해 동적_사건트래킹은 고정_사건트래킹에 비해 좋지 않은 성능을 보인 것으로 판단된다.



〈그림 5〉 고정_ 학습문서_사건트래킹과 동적_ 학습문서_사건트래킹의 오보율 변화

4.2 초기질의 정보필터링과 수정질의 정보필터링의 성능 비교

〈표 4〉는 초기질의를 고정적으로 사용하는 정보필터링과 수정질의를 사용하는 정보필터링의 검색 성능을 보여준다. 누락률은 초기질의_정보필터링이 0.2616으로서 수정질의_정보필터링의 0.1856보다 높았다. 그러나 오보율에 있어서는 초기질의_정보필터링이 0.0264로서 수정질의_정보필터링의 0.1521보다 낮게 나타났다.

재현율에 있어서는 수정질의_정보필터링의 성능이 다소 높게 나타났으나 정확률에 있어서는 초기질의_정보필터링의 성능이 훨씬 좋게 나타났다. 재현율과 정확률을 결합한 F_1 척도에 있어서는 수정질의_정보필터링의 값이 0.1884, 초기질의_정보필터링의 값이 0.5353으로 초기질의_정보필터링의 성능이 더 우수하게 나타났다.

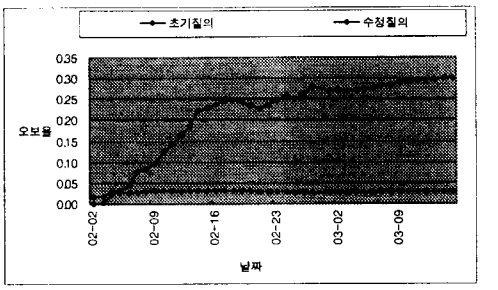
트래킹비용을 기준으로 두 실험의 성능

〈표 4〉 초기질의_정보필터링과 수정질의_정보필터링의 성능 비교

척 도	초기질의	수정질의
누 락 률	0.2616	0.1856
오 보 율	0.0264	0.1521
재 현 율	0.7384	0.8144
정 확 률	0.4462	0.1188
F1	0.5353	0.1884
트래킹비용	0.0086	0.0185

을 비교해보면 초기질의_정보필터링의 트래킹비용은 0.0082, 수정질의_정보필터링은 0.0185로서 초기질의를 사용한 정보필터링의 성능이 수정질의를 사용한 정보필터링보다 더 우수한 것으로 나타났다.

수정질의를 사용한 정보필터링의 성능이 초기질의를 사용한 정보필터링에 비해 떨어지는 원인은 사건트래킹에서와 마찬가지로 필터링 과정에서 적합문서로 잘못 판정된 부적합문서의 비율, 즉 오보율이 높기 때문이라고 할 수 있다. 질의를 수정하는 과정에서 질의에 부적합문서의 내용이 반영됨에 따라 검색되는 문서들 가운데 해당 사건과 관련되지 않은 문서들이 포함되게 되는 것이다. 〈그림 6〉에서 살펴보면, 초기질의_정보필터링의 오보율은 시간의 경과에도 일정한 반면 수정질의_정보필터링의 오보율은 지속적으로 높아지는 경향을 확인할 수 있다.



〈그림 6〉 초기질의_정보필터링과 수정질의_정보필터링의 오보율 변화

4.3 고정 학습문서 사건트래킹과 초기질의 정보필터링의 성능 비교

앞에서 수행한 사건트래킹 기법과 정보필터링 기법에 대한 실험 결과 각 기법에서 보다 나은 성능을 보이는 방법을 선택하여 두 기법의 성능을 비교하였다. 즉 고정 학습문서 집합을 사용한 사건트래킹 기법과 초기질의를 사용한 정보필터링 기법을 선택하여 두 기법의 사건검색 성능을 직접 비교한 결과가 〈표 5〉에 나와 있다.

〈표 5〉를 살펴보면, 누락률에 있어서는 고정_학습문서_사건트래킹이 0.4672, 초기질의_정보필터링이 0.2616으로 정보필터링의 경우 적합문서가 검색에서 누락되는 비율이 더 낮은 것을 알 수 있다. 그러나 오보율을 기준으로 두 기법의 성능을 비교해보면, 고정_학습문서_사건트래킹은 0.0130, 초기질의_정보필터링은 0.0264로서 고정_학습문서_사건트래킹

〈표 5〉 고정_학습문서_사건트래킹과 초기질의_정보필터링의 성능비교

척도	사건트래킹(고정 학습문서 집합)	정보필터링(초기질의)
누락률	0.4672	0.2616
오보율	0.0130	0.0264
재현율	0.5328	0.7384
정확률	0.3977	0.4462
F1	0.4056	0.5353
트래킹비용	0.0141	0.0086

〈표 6〉 사건트래킹과 정보필터링의 사건검색 성능 종합평가

척도	사건트래킹		정보필터링	
	고정 학습문서 집합	동적 학습문서 집합	초기질의	수정질의
누락률	0.4672	0.6705	0.2616	0.1856
오보율	0.0130	0.0611	0.0264	0.1521
F1	0.4056	0.1191	0.5353	0.1884
트래킹비용	0.0141	0.0253	0.0086	0.0185

이 더 좋은 성능을 보인다. 또한 초기질의_정보필터링이 고정_학습문서_사건트래킹에 비해 재현율과 정확률에 있어서도 더 나은 성능을 보이고 있다. F_1 값에서도 초기질의_정보필터링이 0.5353, 고정_학습문서_사건트래킹이 0.4056으로서 마찬가지 경향을 보이고 있다.

사건트래킹과 정보필터링의 성능을 최우선 척도인 트래킹비용을 기준으로 비교해 보면 고정_학습문서_사건트래킹의 트래킹비용은 0.0141로서 초기질의_정보필터링의 0.0086보다 그 값이 더 큰 것을 확인할 수 있다. 따라서 초기질의를 사용한 정보필터링이 고정 학습문서 집합을 사용한 사건트래킹에 비해 모든 평가 척도에 있어서 더 우수한 사건검색 성능을 보였다.

5 종합평가 및 결론

이 논문에서는 기존의 검색대상이 되어 온 주제(topic)와는 달리 특정 사건(event)과 관련된 문서를 지속적으로 검색하여 이용자에게 제공하는 기법에 관해 연구하였다. 연구 대상 기법으로는 kNN 분류기를 사용한 사건트래킹 기법과 질의기반 정보필터링 기법을 선정하여 사건검색 실험을 수행한 후 각 기법의 성능을 비교하였다. 각 기법의 사건검색 성능을 누락률, 오보

율, F_1, 트래킹비용 등 네 가지 주요 척도에 의해 종합적으로 평가한 결과가 〈표 6〉에 나와 있다.

누락률을 기준으로 각 시스템의 성능을 살펴보면 '수정질의_정보필터링 〉 초기질의_정보필터링 〉 고정_학습문서_사건트래킹 〉 동적_학습문서_사건트래킹'의 순으로 우수한 사건검색 성능을 보였다. 즉 정보필터링 기법의 두 가지 방법이 사건트래킹의 두 가지 방법보다 더 나은 성능을 보였으며, 따라서 정보필터링 기법이 사건트래킹 기법보다 누락률의 측면에서 더 나은 성능을 갖는다고 할 수 있다.

오보율을 기준으로 사건검색 성능을 살펴보면 '고정_학습문서_사건트래킹 〉 초기질의_정보필터링 〉 동적_학습문서_사건트래킹 〉 수정질의_정보필터링'의 순으로 좋은 성능을 갖는 것으로 나타났다. 즉 오보율에 있어서는 시스템이 긍정으로 판정한 문서를 각각 트래킹과 필터링 과정에서 활용하는 방법이 낮은 성능을 보였다.

재현율과 정확률을 결합한 단일가 척도인 F_1은 다른 척도들과는 달리 값이 클수록 성능이 우수한 것을 의미한다. F_1을 기준으로 각 시스템의 성능을 비교해 보면 '초기질의_정보필터링 〉 고정_학습문서_사건트래킹 〉 동적_학습문서_사건트래킹 〉 수정질의_정보필터링' 순으로 좋은 성능을 보였다. F_1에 있어서도 오보율의 경우와 마찬가지로 고정된 학습문서집합과 초기질의를 사용하는 방법이 더 우수한 성능을 갖는 경향을 보이고 있다. 이는 동적 학습문서 집합을 사용한 사건트래킹과 수정질의를 사용한 정보필터링의 경우 오보율이 높아짐에 따라 정확률이 낮아지기 때문인 것으로 해석할 수 있다.

마지막으로 누락률 및 오보율과 마찬가지로 값이 작을수록 좋은 성능으로 평가되는 트래킹비용을 기준으로 비교해 보면 '초기질의_정보필터링 〉 고정_학습문서_사건트래킹 〉 수정질의_정보필터링 〉 동적_학습문서_사건트래킹'의 순으로 성능이 우수한 것으로 나타났다.

본 연구의 실험결과를 요약하면 다음과 같다.

첫째, 고정 학습문서 집합을 사용한 사건트래킹이 동적 학습문서 집합을 사용한 사건트래킹보다 더 우수한 사건검색 성능을 보였다.

둘째, 초기질의를 사용하는 정보필터링이 수정질의를 사용하는 정보필터링보다 더 우수한 사건검색 성능을 보였다.

셋째, 고정 학습문서 집합을 사용한 사건트래킹과 초기질의를 사용한 정보필터링의 실험 결과를 중심으로 사건트래킹과 정보필터링 기법을 비교한 결과 정보필터링이 더 좋은 사건검색 성능을 보였다.

결론적으로 특정 사건과 관련된 문서를 검색하여 제공하는 기능에 있어서 kNN 분류기를 사용한 사건트래킹 기법보다 질의기반 정보필터링 기법이 더 효과적인

것으로 나타났다. 이는 긍정 학습문서 집합과 부정 학습문서 집합의 내용이 각각 긍정과 부정으로 반영되는 사건트래킹 기법보다 긍정 학습문서의 내용만을 기반으로 질의를 생성해 사용하는 정보필터링 기법이 사건검색에 더 효과적임을 의미한다. 사건은 주제와는 달리 특정 장소와 시간 등의 구체적인 핵심어를 갖기 때문에 이러한 핵심어들로 구성된 질의를 통해 정보필터링을 수행한다면 해당 사건과는 관계없는 불필요한 용어들이 반영될 수 있는 사건트래킹 기법보다 더 효과적인 사건검색을 수행할 수 있을 것이다. 또한 실험 결과 오보율이 높은 상태에서 시스템이 적합문서로 처리한 실험문서를 긍정 학습문서로 사용하는 것은 각 기법의 사건검색 성능에 부정적인 영향을 끼친다고 할 수 있다.

참 고 문 헌

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

Allan, J., Carbonell, J. 1998. "Topic Detection and Tracking Pilot Study Final Report." *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.*

Allan, J., Lavrenko, V., and Papka, R. 1998. "Event Tracking." *CIIR Technical Report IR-128.*

Ault, T., Yang, Y. 2002. "Information Filtering in TREC-9 and TDT-3: A Comparative Analysis." *Information Retrieval* , 5:159-187.

Carpineto, C. 2001. "An Information-Theoretic Approach to Automatic Query Expansion." *ACM Transactions on Information Systems,* 19(1):1-27.

Hull, D. A., Robertson, S. 2000. "The TREC-8 Filtering Track Final Report". *NISTSP* 500-246:35-56.

Kurt, H. 2001. *On-line New Event Detection and Tracking in Multi-Resource Environment.* M.S.thesis, Bilkent University.

Mostafa, J., Mukhopadhyay, S., Lam, W., and Palakal, M. 1997. "A Multilevel Approach to intelligent Information Filtering: Model, System, and Evaluation." *ACM Transactions on Information Systems,* 15(4): 368-399.

National Institute of Standards and Technology. 2000. Topic Detection and Tracking. [cited 2003.8.1] 〈http://www.nist.gov/speech/tests/tdt/tasks/track.htm〉.

Papka, R. 1999. *Online new event detection, clustering, and tracking.*

Ph.D. diss.,University of Massachusetts Amherst.

Robertson, S. 2002. "Introduction to the Special Issue: Overview of the TREC Routing and Filtering Tasks." *Information Retrieval*, 5:127-137.

Salton. G., Buckly, C. 1990. "Improving retrieval performance by relevance feedback." *Journal of Documentation*, 37(4): 194-214.

Xu, J., Croft, W. B. 1996. "Query Expansion Using Local and Global Document Analysis." *Proceedings of ACM SIGIR International Conference on Research and Development in Information Retrieval*: 4-11.

Yang, Y. 1999. "An evaluation of statistical approaches to text categorization." *Information Retrieval*, 1:69-90.

Yang, Y., Ault, T., Pierce, T., and Lattimer, C.W. 2000. "Improving text categorization methods for event tracking." *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*: 65-72.

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