

효과적인 의견 자질 결합을 위한 실험적 연구

Experimental Study for Effective Combination of Opinion Features

한경수(Kyoung-Soo Han)*

초 록

의견 검색은 사용자의 정보 요구에 주제적으로 연관되면서도 의견이 포함되어 있는 정보를 검색하는 태스크이다. 본 연구는 효과적인 의견 검색을 위해 사용자 정보 요구를 표현하는 방법과 이 요구를 만족시킬만한 여러 의견 자질들을 효과적으로 결합할 수 있는 방법에 대하여 실험을 통해 분석하였다. 본 실험에서는 추론 네트워크 모델을 기본 검색 모델로 사용하였고, Blogs06 컬렉션과 100개의 TREC 토픽에 대해 실험을 수행하였다. 실험 결과, 가상의 '의견' 개념을 설정하여 효과적으로 의견 검색의 정보 요구를 표현할 수 있었으며, 극히 소량의 일반 의견 단어집만을 사용했는데도 동일한 환경에서 기존 모델과 견줄 만한 의견 검색 성능을 달성할 수 있었다.

ABSTRACT

Opinion retrieval is to retrieve items which are relevant to the user information need topically and include opinion about the topic. This paper aims to find a method to represent user information need for effective opinion retrieval and to analyze the combination methods for opinion features through various experiments. The experiments are carried out in the inference network framework using the Blogs06 collection and 100 TREC test topics. The results show that our suggested representation method based on hidden 'opinion' concept is effective, and the compact model with very small opinion lexicon shows the comparable performance to the previous model on the same test data set.

키워드: 의견 검색, 의견 마이닝, 자질 결합, 질의 확장, 의견 단어집
opinion retrieval, opinion mining, feature combination, query expansion, opinion lexicon

* 성결대학교 컴퓨터공학부 전임강사(kshan@sungkyul.ac.kr)

■ 논문접수일자: 2010년 8월 14일 ■ 최초심사일자: 2010년 8월 21일 ■ 게재확정일자: 2010년 8월 27일
■ 정보관리학회지, 27(3): 227-239, 2010. [DOI:10.3743/KOSIM.2010.27.3.227]

1. 서론

최근 인터넷 사용자들은 블로그(blog), 게시판, 뉴스 댓글 등을 통해 다양한 방식으로 의견을 표출하고 있으며, 이 의견 정보들은 다른 인터넷 사용자나 기업에게 유용한 정보가 되고 있다. 특히 블로그는 기존 언론이 담지 못한 다양한 의견을 제시함으로써 사회 여론 형성에 큰 역할을 하고 있다. 더불어 트위터나 페이스북 같은 소셜 네트워크 서비스가 등장하면서 인터넷 공간은 단순 사실(fact) 보다는 감정이나 의견 등을 교류하는 공간으로 바뀌어가고 있다. 이런 상황에서 단순 사실에 대한 검색을 넘어 의견을 찾는 의견 검색(opinion retrieval)의 필요성이 대두되었으며, 그 중요성은 날로 커지고 있다.

의견 검색을 위해서는 기존의 사실을 검색하는 주제 검색(topical retrieval)의 기능뿐만 아니라 추가적으로 의견을 탐지하는 기능이 필요하다. 예를 들어, 사용자가 "macbook pro"라는 질의어를 입력했다고 했을 때, 기존의 일반 주제 검색 시스템에서는 "macbook pro"라는 제품과 연관된 모든 정보를 검색하는 것이 목표이다. 이때 검색되는 정보는 "macbook pro"에 대한 사실 위주의 정보들일 것이다. 그러나 의견 검색 시스템의 목표는 "macbook pro"라는 제품에 대해 사용자들이 어떤 반응과 의견을 보이는지를 알아보기 위해 의견이 포함되어 있는 정보를 검색하는 것이 목표이다. 이때 검색되는 의견 정보는 당연히 "macbook pro"와 연관된 적합 정보일 것이다. 따라서, 의견 검색은 주어진 사용자 질의에 적합하면서도 그 질의에 대한 의견을 포함하는 정보를 검색하는 문제이다.

의견 검색과 관련하여 정보 검색, 텍스트 마이닝, 자연어처리 분야에서 의견 검색, 의견 마이닝(opinion mining), 감성 분석(sentiment analysis) 등의 이름으로 다양한 연구가 진행 중이다. 특히 미국 NIST(National Institute of Standards and Technology)에서 주관하는 TREC(Text REtrieval Conference)에서 2006년도부터 블로그 글에서 특정 주제에 대한 의견을 검색하는 블로그 트랙(blog track)을 시작하면서 의견 검색에 대한 연구가 활발히 진행되고 있다. 그러나 블로그 트랙에 제출된 많은 의견 검색 시스템들의 검색 결과가 일반 사실 기반의 주제 검색 성능을 향상시키지 못하거나 향상시키더라도 향상 폭이 매우 미미한 것으로 나타났다(Ounis, Macdonald, and Soboroff 2008; Macdonald, Ounis, and Soboroff 2009).

이 문제를 해결하기 위한 일환으로 본 연구는 의견 검색의 사용자 정보 요구 표현과 자질(feature)들의 효과적인 결합 방법에 대해 실험을 통해 분석한다. 예를 들어, "macbook pro"라는 상품에 대한 블로거들의 의견을 검색하고자 한다면, 대개의 경우 사용자는 의견 검색 시스템에 단지 "macbook pro"라는 검색 대상 주제어만을 입력한다. 입력된 주제어만으로 검색을 수행한다면 일반 주제 검색과 다를 바가 없다. 따라서 입력된 주제어 이외에 숨겨져 있는 '의견에 대한 검색'이라는 사용자 정보 요구를 표현하는 방법과 이를 만족시킬만한 자질들을 효과적으로 결합하는 방법에 대한 연구가 필요하다.

본 논문에서는 '의견에 대한 검색'을 표현하기 위해 '의견'이라는 가상의 개념을 설정하고 이에 대한 여러 자질들을 '의견' 개념으로 통합하는 방법을 제안한다. 자질 통합을 위해서는

여러 자질 정보를 효과적으로 결합할 수 있다고 알려진 추론 네트워크 모델을 이용하였다.

본 논문의 구성은 다음과 같다. 2절에서는 일반적인 의견 검색시스템에 대한 관련 연구를 살펴보고, 3절에서 본 실험의 틀로 사용하는 추론 네트워크 모델에 대해 설명한다. 4절에서는 실험 데이터를 비롯한 실험 환경을 설명하고, 5절에서 실험 내용을 설명하고 그 결과를 분석한다. 마지막 6절에서 본 연구의 결론과 향후 연구에 대해 논한다.

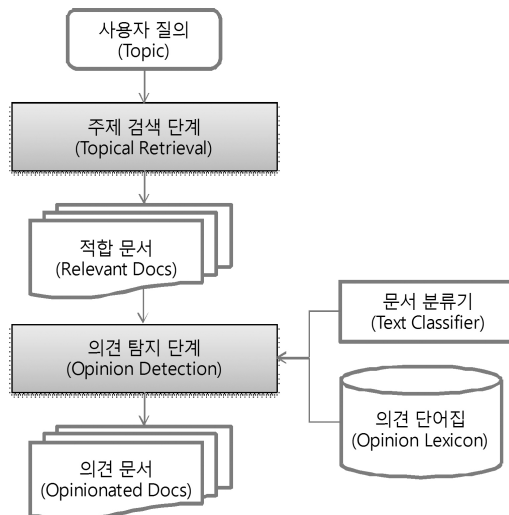
2. 관련 연구

본 논문에서 다루는 의견 검색은 블로그 문서들을 대상으로 사용자 질의에 연관되면서도 의견을 포함하고 있는 블로그 문서를 검색하는 것을 목적으로 한다. 이런 측면에서 기존의 일반적인 주제 검색의 틀에서 대상 문서 도메인

만 블로그로 한정되어 있는 블로그 검색과는 차이가 있다.

의견 검색에 대해 가장 활발히 연구가 진행되고 있는 TREC의 블로그 트랙을 위주로 관련 연구를 분석한다. TREC 참여 시스템 대부분은 <그림 1>과 같이 2단계로 구성되어 있다. 기존의 일반 검색시스템과 같이 주어진 주제(topic)에 적합한 문서를 검색하는 주제 검색 단계를 거친 후, 의견 탐지를 위한 여러 자질을 기반으로 문서 순위를 재조정하는 의견 탐지(opinion detection) 단계를 거쳐 주제에 적합하면서도 의견이 포함된 문서를 검색한다.

의견 탐지를 위해 의견 검색의 대부분 연구에서 의견을 표현하는 단어 및 구절들을 모아 놓은 의견 단어집(opinion lexicon)을 활용한다. Turney and Littman(2003)은 소량의 초기 의견 단어와의 통계적인 단어 연관도를 측정함으로써 의견 단어집을 자동 구축하였다. Yang 등(2007)은 의견 단어 외에 의견 연어(colloca-



<그림 1> 일반적인 2단계 의견 검색시스템 구조

tion)를 활용하였는데, 의견 표현 블로그에서 'I'나 'you'같은 대명사가 자주 사용된다는 사실에 기반하여, 학습 데이터로부터 'I believe', 'my assessment', 'good for you' 등의 연어를 추출하였다. 또 주관적인 형용사와 동사가 집중적으로 등장하는 문서는 의견을 표현하고 있을 것이라는 가정 하에 형용사-동사 연어도 추출하여 이용하였다. Vechtomova(2007)은 주관적인 형용사를 수동으로 구축하였다. 또한 Kim과 Hovy(2005)는 수동으로 구축된 소규모 의견 단어를 WordNet의 동의어(synonym)와 반의어(antonym)를 이용하여 확장하는 방식으로 의견 어휘집을 구축하였고, Zhang과 Zhang(2006)은 SentiWordNet에 기반하여 주관적인 형용사를 추출하였다. 한편, 의견 문서 및 비의견 문서에서의 단어 출현 통계를 이용하여 의견 단어에 가중치를 부여하기도 하였다(이승욱, 송영인, 임해창 2008; He et al. 2008). 의견 어휘집 구축에 대한 한국어 연구에서는 문장 단위로 긍/부정 정보가 태깅된 리뷰 문장으로부터 의견 사전을 추출하는 연구가 주류를 이루었다(남상협 등 2008).

의견 탐지를 위해 학습 데이터를 구축하여 문서 분류 문제로 해결하는 방법론들도 있었다. 학습을 시키기 위해서는 주관적인 문서 및 객관적인 문서가 필요하다. 기존 연구들에서는 특정 주제에 대한 리뷰 사이트에서 주관적인 문서를, 위키피디아(Wikipedia)나 신문기사로부터 객관적인 문서를 수집하여 학습 문서로 활용하였다. 이 학습문서를 바탕으로 SVM(Support Vector Machine)(Jia, Yu, and Zhang 2008), 지수회귀 모형(logistic regression model)(Zhang and Zhang 2006), 잠재의미색인(LSI)(이지혜, 정

영미 2009) 등의 분류기를 사용하였다.

의견 탐지 단계에서 발견된 의견이 해당 주제와 연관되지 않은 것일 수도 있으므로 주제와 연관된 의견을 담고 있는 문서를 찾아야 한다. 이를 위해 의견과 주제 사이의 관계를 고려해야 하는데, 기존 연구들은 주제어와 의견 단어가 출현한 위치를 기반으로 근접하여 등장한 경우 해당 의견은 주제와 연관된 것으로 간주하였다(Vechtomova 2007; Zhou, Joshi, and Bayrak 2007; Jia, Yu, and Zhang 2008; 윤홍준, 김한준 2010). 특히 Gerani 등(2010)은 근접 출현 정보를 다른 자질들과 효과적으로 결합할 수 있는 의견 검색 모델을 제안하여 좋은 성능을 보였다.

의견 문서의 순위 부여를 위해 주제 검색 단계의 적합성 점수와 주제 탐지 단계의 의견 점수를 결합하여야 한다. 기존 연구들에서는 선형 결합과 로그 결합 등이 제안되었으며(He et al. 2008; Jia, Yu, and Zhang 2008), 단순한 점수 결합의 문제점을 극복하고자 이론적인 기반 위에 모델 차원에서 통합하려는 시도가 있었다(Zhang and Ye 2008). 긍/부정에 대한 의견 극성(polarity)의 엔트로피를 순위 부여에 적용한 연구도 있었다(윤홍준, 김한준 2010).

본 논문에서 초점을 맞추는 의견 검색의 정보 요구와 자질 결합에 관한 연구도 있었다. Huang과 Croft(2009)는 의견 검색의 정보 요구를 표현하기 위한 의견 적합성 모델(opinion relevance model)을 제안하였다. 즉, 주제어만으로 구성된 질의에 추가할 의견 단어를 선정하는 방법을 제안하였는데, 다양한 말뭉치를 활용하는 질의 독립적인 것과 의사 적합성 피드백(pseudo relevance feedback)을 이용한 질의 의존적인 방

법이 그것이다. 본 논문은 이 연구와 궤를 같이 하여 추가적인 실험을 통해 그 결과를 분석하고자 한다.

3. 추론 네트워크 모델

3.1 추론 네트워크 모델 개요

이 절에서는 본 논문의 실험에서 사용하는 추론 네트워크 모델(inference network model)에 대해 소개한다. 추론 네트워크는 방향성 비순환 그래프(directed acyclic graph: DAG)로서, 노드는 사건(event)을 나타내고 간선은 사건 사이의 확률적 의존관계를 표현한다. 검색 모델로 사용될 때 노드는 특정 문서의 관측과 같은 사건, 특정 자질, 혹은 자질들의 조합 등을 나타낸다. 추론 네트워크 모델은 검색을 위한 여러 자질 정보들을 효과적으로 결합할 수 있다는 장점이 있다(Croft, Metzler, and Strohman 2009).

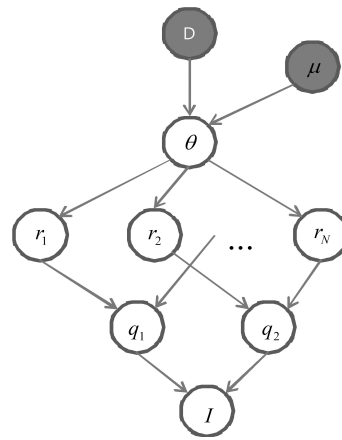
질의에 적합하면서도 의견이 포함된 문서를 검색하는 의견 검색에서는 주제와의 적합성을 평가하는 주제 자질과 의견의 포함 정도를 가늠하는 의견 자질 등 서로 이질적인 다양한 자질을 결합해야 한다. 따라서 본 연구에서는 다양한 자질을 효과적으로 결합할 수 있는 추론 네트워크 모델을 이용한다.

〈그림 2〉는 문서 하나에 대한 일반적인 추론 네트워크 모델의 예를 보인 것이다. 표현 노드(representation node) r_i 는 문서에 대한 자질에 해당하며, 확률값은 파라미터 μ 를 사용해 추정되는 언어모델 θ 로부터 계산된다. 질의 노

드(query node) q_i 는 표현 노드나 다른 질의 노드로 부터의 자질들을 결합하는데 사용된다. 결국 추론 네트워크는 문서 D 와 파라미터 μ 가 주어졌을 때 사용자 정보 요구 I 가 만족될 확률 $P(I|D, \mu)$ 를 계산하게 된다. 이를 위해 표현 노드들은 다음과 같이 〈수식 1〉에 따라 확률이 계산된다.

$$P(r_i|D, \mu) = \frac{f_{i,D} + \mu P(r_i|C)}{|D| + \mu} \quad (1)$$

여기서 $f_{i,D}$ 는 자질 r_i 가 문서 D 에서 출현한 회수, $P(r_i|C)$ 는 전체 컬렉션에서의 r_i 확률, μ 는 Dirichlet 평탄화(smoothing) 파라미터이다.



〈그림 2〉 추론 네트워크 모델의 예

3.2 주요 연산자

본 연구에서는 자질들을 결합하는데 추론 네트워크 모델을 이용하며, 실험에는 Lemur Toolkit에 포함된 Indri 검색 엔진(Strohman et al. 2005)을 이용하였다. Indri 검색 엔진은 최근 검

색 연구들에서 가장 좋은 성능을 보이는 언어 모델(language model)에 기반하여 계산된 자질값들을 효과적으로 결합할 수 있으며, 다양한 형태의 질의를 실험해 볼 수 있어 여러 검색 실험들에서 활용되고 있다. Indri에서 질의를 표현 하는데 사용되는 질의 언어 중 본 실험과 연관된 주요 연산자에 대해서만 간략히 살펴본다.

- **#syn**: 동의어(synonym) 리스트를 표현 하는데 사용된다. 예를 들면, #syn(car automobile)과 같이 표현하면 마치 '자동차'라는 새로운 개념이 존재하는 것처럼 간주하여, 단어 빈도는 'car'와 'automobile' 각 단어 빈도를 합하여 계산하고, 문서 빈도는 두 단어가 출현한 문서의 합집합의 원소 개수로 계산된다.
- **#odx**: 순서가 유지된 채 x 단어 이내에 출현한 단어들을 하나의 자질로 고려한다. 예를 들어, #od3(blue car)는 단어 'car' 앞쪽 3개 이내의 단어에 'blue'가 출현한 경우 자질로 고려된다.
- **#uwx**: 순서와 상관없이 x 단어 이내에 출현한 단어들을 하나의 자질로 고려한다. 예를 들어, #uw3(blue car)는 단어 'car' 앞뒤 3 단어 이내에 'blue'가 출현한 경우 자질로 고려된다.
- **#combine**: n 개의 자질을 결합하는 연산자로서 <수식 2>와 같이 각 자질 확률의 기하평균을 계산한다. 이때 자질 r_i 에 대한 확률 p_i 는 앞서 설명한 <수식 1>에 의해 계산된다.

$$P_{combine} = \prod_i p_i^{1/n} \quad (2)$$

- **#weight**: #combine과 동일하나 각 자질 확률에 가중치 w_i 를 부여하여 <수식 3>과 같이 자질들을 결합한다.

$$P_{weight} = \prod_i p_i^{w_i / \sum_j w_j} \quad (3)$$

- **#wsum**: n 개의 자질 확률에 가중치 w_i 를 부여하여 <수식 4>와 같이 가중치 산술평균으로 자질을 결합하는 연산자이다. 이때 가중치 w_i 가 모든 자질에 동일하게 부여된다면 일반 산술평균이 계산된다.

$$P_{wsum} = \frac{\sum_i w_i p_i}{\sum_i w_i} \quad (4)$$

- **#max**: <수식 5>와 같이 n 개의 자질 확률 중 최대값을 사용한다.

$$P_{max} = \max_i p_i \quad (5)$$

4. 실험 환경

4.1 실험 데이터

4.1.1 문서 집합

TREC 블로그 트랙에서 시스템 간의 성능 평가에 사용했던 Blogs06 컬렉션을 본 논문의 실험에 사용하였다. 이 컬렉션은 2005년 12월 6일부터 2006년 2월 21일까지 11주 동안 수집된 블로그 데이터이다(Macdonald and Ounis 2006). Blogs06 컬렉션은 총 크기가 148GB인데, 블로그 자체를 의미하는 XML 피드(38.6GB), 개개 블로그 문서와 커멘트를 의미하는 고유링크

(permlink) 문서(88.8GB), 블로그의 첫페이지에 해당하는 HTML 홈페이지(28.8GB) 등 3가지 요소로 구성된다. 의견 검색 태스크에서는 고유링크 문서가 검색 대상이 되는데, 총 3,215,171개의 문서가 포함되어 있다. 2009년도 블로그 트랙에서는 새롭게 구축된 Blogs08 컬렉션을 사용하는데, 데이터 크기가 2TB를 넘는 대용량이다. Blogs08 컬렉션은 대용량으로 인해 실험에 어려움이 있고, 기존 연구와의 실험 결과 비교를 위해 본 실험에서는 Blogs06 컬렉션을 사용한다.

4.1.2 질의 집합

검색 질의는 2006년도부터 2008년도까지 3년 동안 TREC에서 사용했던 총 150개(851번~950번, 1001번~1050번)의 토픽(topic)을 사용한다. 하나의 토픽은 <그림 3>과 같이 title, description, narrative 3가지로 구성되는데 본 실험에서는 title만으로 질의를 구성하였다. Huang and Croft(2009) 연구와의 실험 결과를 비교할 수 있도록 실험 환경을 맞추기 위해 851번~900번 50개의 토픽은 학습 및 파라미터 조정용으로, 901번 이후의 100개의 토픽은 실험용으로 사용하였다.

4.1.3 의견 문서 집합

성능을 평가하기 위해서는 각 질의에 대한 의견 문서 집합이 필요한데 2006년부터 2008년까지의 TREC 평가에서 사용한 의견 문서 집합을 사용한다. 이 의견 문서 집합은 각 질의에 대해 TREC 참가 시스템들이 제출한 1,000개의 검색 결과 문서를 평가자들이 수동으로 질의에 대한 적합성 여부, 질의에 대한 의견 포함 여부, 의견의 긍/부정 여부 등을 판정한 결과이다.

4.2 의견 단어집

본 논문에서는 기존 연구들에서 보고된 소량의 의견 단어만을 사용한다. 실험에 사용된 의견 단어의 종류는 <표 1>과 같다. 이 중 G2, G4, G14의 의견 단어들은 본 실험에서 사용하는 실험 데이터와는 무관한 일반적인 의견 단어들이며, T5 의견 단어집은 Huang and Croft(2009)가 본 실험에서 사용하는 Blogs06 컬렉션에 대해 학습 질의에 해당하는 851번~900번 토픽을 사용하여 추출한 의견 단어이다.

```

<top>
<num> Number: 856 </num>
<title> macbook pro </title>
<desc> Description:
What has been the reaction to the Macbook Pro laptop computer?
</desc>
<narr> Narrative:
General statements of liking or disliking the Macbook Pro are relevant. Value comparisons to earlier versions of Macintosh laptops or to other companies laptops are relevant. Product reviews are relevant if they contain opinions. Speculation about unreleased laptops is not relevant.
</narr>
</top>
    
```

<그림 3> TREC 토픽 예: "macbook pro"

〈표 1〉 실험에 사용된 의견 단어의 종류

구분	내용	구성 단어
G2	Turney and Littman(2003)이 사용한 초기 일반 의견 단어 중 2개	good, bad
G4	Huang and Croft(2009)가 실험에 사용한 일반 의견 단어 4개	good, bad, nice, poor
G14	Turney(2003)가 사용한 초기 일반 의견 단어 총 14개	good, nice, excellent, positive, unfortunate, correct, superior, bad, nasty, poor, negative, unfortunate, wrong, inferior
T5	Huang and Croft(2009)가 Blogs06 학습데이터에서 추출한 의견 단어 5개	even, like, know, too, good

4.3 평가 척도

본 논문에서는 효율성(efficiency)보다는 효과성(effectiveness)에 초점을 맞춰 성능 평가를 수행한다. 성능 평가 척도로는 TREC과 기존 연구들에서 주로 사용하는 평균 정확률(MAP), R-정확률(R-prec), 10위까지의 정확률(P@10) 등을 사용하였다.

5. 실험 및 평가

5.1 의견 자질의 결합 방식

주제어 자질 t_i 와 의견 자질 o_i 를 결합하기 위해 사용한 질의 q_{opin} 은 〈수식 6〉과 같다.

$$q_{opin} = \#combine(t_1 t_2 \dots t_m \#op_{opin}(o_1 o_2 \dots o_n)) \quad (6)$$

여기서 $\#op_{opin}$ 은 의견 자질에 대해 어떠한 연산을 수행하여 주제어 자질과 결합할 것인지, 즉 의견 자질의 결합 방식을 지정하는 연산자이다. 〈표 2〉는 의견 자질의 결합 방식에 따라 의견 검색 시스템의 성능을 비교한 것이다. 이 실험에는 T5 의견 단어집이 사용되었다. 〈표 2〉에서 Baseline은 의견 단어 $\{o_1, o_2, \dots, o_n\}$ 는 사용하지 않고 주제어 $\{t_1, t_2, \dots, t_m\}$ 만으로 검색한 결과이다. No는 의견 단어에 대해 어떠한 연산도 수행하지 않고 기존 질의어들과 그대로 결합한 방법이다. 즉, 〈수식 7〉이 질의로 사용된 결과이다.

$$q_{opin} = \#combine(t_1 t_2 \dots t_m o_1 o_2 \dots o_n) \quad (7)$$

Sum은 〈수식 6〉의 $\#op_{opin}$ 에 3,2절에서 설명

〈표 2〉 의견 자질의 결합 방식에 따른 성능 비교

결합 방식	Opin_ret	MAP	R-prec	P@10
Baseline	10483	0.2629	0.3184	0.4740
No	10521	0.2729	0.3159	0.5210
Sum	10485	0.2630	0.3185	0.4740
Max	10483	0.2629	0.3184	0.4740
Syn	11144	0.3039	0.3481	0.5360

한 #wsum 연산자를 사용하여 각 자질의 확률에 대한 산술평균으로 결합하는 방법을, Max는 #max 연산자를 사용하여 각 자질의 확률 중 최대값을 사용하여 결합하는 방법을 뜻한다. Sum에서 가중치는 모두 동일하게 부여하여 일반 산술평균을 계산하였다. Syn은 #*op_{opin}*에 #syn 연산자를 이용하여 의견 단어들을 동의어인 것처럼 간주하는 방법을 뜻한다. Opin_ret은 검색된 의견 문서의 개수를 의미한다.

〈표 2〉에서 보듯이 Sum이나 Max 방법은 No보다도 성능이 낮았으나, Syn 방법은 성능향상을 보였다. 즉, 의견 자질을 개별적으로 고려하여 결합하는 것보다 하나의 개념으로 통합하여 결합하는 것이 효과적이었다. 의견 단어들이 단어 의미적으로 봤을 때는 반드시 동의어라고 말할 수는 없으나, 의견 검색 관점에서는 ‘의견’이라는 하나의 개념을 표현하는 동의어로 간주할 수 있다. Syn 결합 방법을 통해 여러 의견 단어들의 정보를 결합한 가상의 ‘의견’ 개념을 설정하여 주제어 자질과 의견 자질을 효과적으로 결합함으로써 성능 향상이 가능했던 것으로 추정된다. 따라서 본 논문에서 제안한 가상의 ‘의견’ 개념을 통해 의견 자질을 결합하는 방법은 의견 검색의 사용자 정보 요구를 효과적으로 표현할 수 있는 방법이라 하겠다.

5.2 의견 단어별 성능 비교

의견 자질의 종류나 그 크기에 따라 성능이 어떻게 달라지는지를 알아보기 위한 실험을 수행하였다. 〈표 3〉은 No 결합 방식으로 결합했을 때의 성능을, 〈표 4〉는 Syn 결합 방식으로 결합했을 때의 성능을 나타낸다. G2 U T5는 단어집 G2와 T5를 합하여 사용한 경우를 뜻한다.

〈표 3〉에서 보듯이 기존 질의에 포함되어 있는 주제어와 의견 단어를 그대로 결합한 경우에는 의견 단어가 늘어남에 따라 성능이 큰 폭으로 저하되기도 하였다. 2개의 의견 단어가 사용된 G2 단어집을 사용한 경우가 가장 좋은 성능을 보였다.

반면 Syn 결합 방식으로 결합한 경우에는 〈표 4〉에서 보듯이 의견 단어가 추가되어 단어집 크기가 커지더라도 높은 성능을 보일 수 있었다. 이와 관련하여 Huang과 Croft(2009)는 본 실험과 동일한 환경에서 의견 단어 개수에 따라 검색 성능이 저하되기도 하여 의견 단어가 성능 향상에 항상 도움이 되는 것은 아니라고 보고하였다. 하지만 이것은 본 실험 결과와 상충되는 것이다. 이는 의견 자질의 결합 방식에 그 원인이 있는 것으로 추정되며, 본 논문에서 제안하는 Syn 방법으로 결합했을 때에는 의견

〈표 3〉 의견 단어의 종류에 따른 성능 비교(결합 방식: No)

의견 단어집	Opin_ret	MAP	R-prec	P@10
G2	9810	0.2484	0.2971	0.4800
G4	7780	0.1698	0.2261	0.3720
G14	1675	0.0106	0.0276	0.0550
G2 U T5	9611	0.2380	0.2815	0.4740
G4 U T5	6985	0.1532	0.2075	0.3620
G14 U T5	1522	0.0088	0.0242	0.0530

〈표 4〉 의견 단어의 종류에 따른 성능 비교(결합 방식: Syn)

의견 단어집	Opin_ret	MAP	R-prec	P@10
G2	10617	0.2816	0.3263	0.5280
G4	10696	0.2843	0.3290	0.5260
G14	10910	0.2886	0.3359	0.5280
G2 ∪ T5	11116	0.3029	0.3473	0.5340
G4 ∪ T5	11126	0.3035	0.3484	0.5350
G14 ∪ T5	11174	0.3038	0.3496	0.5380

단어가 지속적으로 성능 향상에 기여하였다. 본 실험 결과로 미루어 보건데 의견 단어를 좀 더 정교하게 결합한다면 의견 단어는 성능 향상에 큰 도움이 된다고 할 수 있겠다.

5.3 주제와 의견 사이의 관련성

효과적인 의견 검색을 위해서는 문서에 나타난 의견이 검색 대상 주제에 대한 의견인지를 판단할 수 있는 근거가 필요하다. 전술한 바와 같이 기존 연구들에서는 이 관계를 주제를 표현한 질의어와 의견 단어가 근접하여 출현하면 서로 관계가 있다고 판단하였다. 본 실험에서는 근접 출현 정보를 추론 네트워크의 자질로 추가하기 위해 주제어와 의견 단어가 특정 윈도우 크기 이내에 출현하면 근접 출현으로 간주하였다. 근접 출현 정보에 대한 질의 q_{prox} 는 〈수식 8〉과 같다.

$$q_{prox} = \#combine(\#op_{prox}(t_1 \#sym(o_1 o_2 \dots o_n)) \#op_{prox}(t_2 \#sym(o_1 o_2 \dots o_n)) \dots \#op_{prox}(t_m \#sym(o_1 o_2 \dots o_n))) \quad (8)$$

여기서 $\#op_{prox}$ 는 근접 출현 여부를 판정하기 위한 윈도우 설정 방법과 크기를 결정하는 연산자를 의미한다. 〈표 5〉는 윈도우 크기에 따른 검색 성능을 보여준다. 이 실험에는 T5 의견 단어집과 Syn 결합 방법이 사용되었다. 표에서 U_x 는 〈수식 8〉의 $\#op_{prox}$ 에 $\#uw_x$ 연산자가 사용된 경우이며, U_∞ 는 $\#uw$ 로 계산하되 윈도우 크기에 제한을 두지 않은 경우이다. O_∞ 는 $\#op_{prox}$ 로 $\#od$ 연산자를 사용하여 윈도우 크기에 제한을 두지 않되 출현 순서는 의견 단어보다 주제어가 먼저 출현한 경우만 고려하도록 제한한 결과이다.

〈표 5〉 근접 출현 윈도우 크기별 성능 비교

윈도우 크기	Opin_ret	MAP	R-prec	P@10
U30	6990	0.1881	0.2665	0.5370
U60	8290	0.2382	0.2998	0.5530
U120	9276	0.2692	0.3187	0.5810
U240	9723	0.2785	0.3304	0.5940
U_∞	9879	0.2819	0.3333	0.5980
O_∞	9579	0.2659	0.3132	0.5880

〈표 5〉에서 보듯이 기대와는 달리 윈도우 크기가 커짐에 따라 성능이 향상 되었다. 윈도우 크기와 상관없이 주제어와 의견 단어가 같이 출현했다는 사실이 의견 문서 판정에 기여한다는 의미이다. 이 결과는 본 실험에 사용된 의견 단어량이 극히 적어 의견 단어의 출현 정보를 가능한 한 많이 고려했을 때 성능이 향상되는 것으로 추정된다.

한편, 윈도우 크기를 무제한으로 했을 때, 주제가 의견 단어 보다 먼저 출현하도록 제약한 유순서 무제한 방법(O_∞)이 순서를 고려하지 않은 무순서 무제한 방법(U_∞)보다 성능이 더 낮았다. 즉, 의견을 표현할 때 반드시 의견 대상이 되는 주제가 먼저 등장하고 의견어가 후위에 위치하지는 않는다는 것이다. 이는 주제를 서로 다른 단어로 다양하게 표현할 수 있기 때문으로 생각되며, 보다 높은 성능을 위해서는 조응어(anaphora)나 지시 표현(referring expression) 등에 대한 처리가 필요하겠다.

5.4 모든 자질의 결합

주제어에 대한 자질, 의견 단어에 대한 자질, 질의어와 의견 단어의 근접 출현에 대한 자질을 모두 결합하기 위한 질의 q 는 〈수식 9〉와 같다.

$$q = \text{weight} \left(\begin{matrix} \alpha \text{ \#combine}(t_1 t_2 \dots t_m) \\ \beta \text{ \#syn}(o_1 o_2 \dots o_n) \\ \gamma \text{ } q_{prox} \end{matrix} \right) \quad (9)$$

여기서 α, β, γ 는 각 자질에 대한 가중치이며 이 가중치는 학습용 질의를 이용하여 적절한 값으로 조정되었다. 〈표 6〉은 모든 자질이 결합된

최종 검색 시스템의 성능을 보여준다. 이 실험에는 G14 U T5 의견 단어집, Syn 결합 방식, 근접 출현 윈도우 크기 U240이 사용되었다. 표에서 Huang and Croft(2009) 결합 모델의 실험 결과는 본 실험 데이터와 동일한 Blogs06 컬렉션에 대해 해당 연구에서 밝힌 성능 수치이다.

〈표 6〉 모든 자질을 결합한 최종 성능

구분	MAP	R-prec	P@10
최종 결합 모델	0.3170	0.3561	0.5780
Huang and Croft(2009) 결합 모델	0.3147	0.3546	0.5640

〈표 6〉에서 보듯이 소수의 의견 단어만을 이용했는데도 본 논문의 최종 결합 모델이 Huang and Croft 모델과 견줄 만한 성능을 보인다. 그러므로 효과적인 의견 검색을 위해서는 다양한 의견 자질도 필요하겠지만, 그 자질들을 결합하는 정교한 결합 방법이 필요하다고 하겠다.

6. 결론

본 연구는 의견 검색의 사용자 정보 요구를 표현하기 위해 소량의 의견 단어집을 이용하는 방법을 제안하였고, 의견 자질들을 효과적으로 결합하기 위한 방법을 실험을 통해 분석하였다. 의견 검색의 사용자 정보 요구를 표현하기 위해 ‘의견’이라는 가상의 개념을 설정하고 각 의견 단어 자질들의 정보를 ‘의견’ 개념으로 통합하여 표현하는 것이 효과적이었다. 즉, 각 의견 단어 자질을 개별적으로 주제어 자질과 결합하는 것보다는 의견 단어 자질을 ‘의견’ 개념으로

통합한 후 주제어 자질과 결합하는 것이 보다 높은 성능을 보였다. 또한, 소량의 일반적인 의견 단어라도 기존 주제어와 정교하게 결합할 경우 의견 검색 성능 향상에 크게 기여함을 알 수 있었다. 의견 단어량이 증가하더라도 지속적으로 성능을 향상시킬 수 있었으며, 기존 모델과 비교해서도 동일한 실험 데이터에서 견줄 만한 성능을 보였다.

본 연구 결과를 바탕으로 향후 대량의 의견 단어집이나 의견 문서 말뭉치도 활용할 수 있도

록 모델을 일반화시켜 확장시키는 연구와 의견 단어가 늘어남에 따라 주제와 의견 사이의 연관성을 측정하기 위한 보다 효과적인 방법에 대한 추가 연구가 필요하다. 또한 의견이 시간적인 차이를 두고 블로그에 등록될 수 있으므로 주제 탐지 및 추적(topic detection & tracking)이나, 유사 이용자들이 의견을 자주 등록하므로 협업여과(cooperative filtering), 연관규칙(association rule) 등의 적용도 고려해볼 수 있을 것이다.

참 고 문 헌

- 남상협, 나승훈, 이예하, 이용훈, 김준기, 이종혁. 2008. 의견 어구 추출을 위한 생성 모델과 분류 모델을 결합한 부분 지도 학습 방법. 『한국정보과학회 2008 종합학술대회 논문집』, 35(1C): 268-273.
- 윤홍준, 김한준. 2010. 오피니언 마이닝 기술을 이용한 효율적 상품평 검색 기법. 『정보과학회논문지: 컴퓨팅의 실제 및 레터』, 16(2): 222-226.
- 이승욱, 송영인, 임해창. 2008. 혼합 방식에 기반한 의견 문서 검색 시스템. 『정보관리학회지』, 25(4): 115-129.
- 이지혜, 정영미. 2009. 지도적 잠재의미색인(LSI) 기법을 이용한 의견 문서 자동 분류에 관한 실험적 연구. 『정보관리학회지』, 26(3): 451-462.
- Croft, W. B., D. Metzler, and T. Strohan. 2009. *Search Engines: Information Retrieval in Practice*. Pearson, 271-283.
- He, B., C. Macdonald, J. He, and L. Ounis. 2008. "An Effective Statistical Approach to Blog Post Opinion Retrieval." *Proceeding of the 17th ACM Conference on Information and Knowledge Management(CIKM-2008)*, 1063-1072.
- Huang, X. and W. B. Croft. 2009. "A Unified Relevance Model for Opinion Retrieval." *Proceedings of the 18th ACM Conference on Information and Knowledge Management(CIKM-2009)*, 947-956.
- Jia, L., C. Yu, and W. Zhang. 2008. "UIC at TREC 2008 Blog Track." *Proceedings of the 17th Text Retrieval Conference (TREC-2008)*.
- Gerani, S., M. J. Carman, F. Crestani. 2010. "Proximity-Based Opinion Retrieval."

- Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR-2010)*, 403-410.
- Kim, S.-M. and E. Hovy. 2005. "Automatic Detection of Opinion Bearing Words and Sentences." *Proceedings of the 2nd International Joint Conference on Natural Language Processing(IJCNLP-2005)*, 61-66.
- Macdonald, C. and L. Ounis. 2006. "The TREC Blogs06 Collection: Creating and Analysing a Blog Test Collection." *DCS Technical Report TR-2006-224*, University of Glasgow.
- Macdonald, C., I. Ounis, and I. Soboroff. 2009. "Overview of the TREC-2009 Blog Track." *Proceedings of the 18th Text Retrieval Conference(TREC-2009)*.
- Ounis, I., C. Macdonald, and I. Soboroff. 2008. "Overview of the TREC-2008 Blog Track." *Proceedings of the 17th Text Retrieval Conference(TREC-2008)*.
- Strohman, T., Metzler, D., Turtle, H., and Croft, W. B. 2005. "Indri: A Language-Model based Search Engine for Complex Queries(Extended Version)." *CIIR Technical Report*.
- "Text Retrieval Conference(TREC)." [online]. [cited 2010.6.15]. <<http://trec.nist.gov>>.
- "The Lemur Project." [online]. [cited 2010.6.15]. <<http://www.lemurproject.org>>.
- "The BLOGS08 test collection." [online]. [cited 2010.6.15]. <http://ir.dcs.gla.ac.uk/test_collections/blogs08info.html>.
- Turney, P. D. and M. L. Littman. 2003. "Measuring Praise and Criticism: Inference of Semantic Orientation from Association." *ACM Transactions on Information Systems*, 21(4): 315-346.
- Vechtomova, O. 2007. "Using Subjective Adjectives in Opinion Retrieval from Blogs." *Proceedings of the 16th Text Retrieval Conference(TREC-2007)*.
- Yang, K., N. Yu, A. Valerio, H. Zhang, and W. Ke. 2007. "Fusion Approach to Finding Opinions in Blogosphere." *Proceedings of the 1st International Conference on Weblogs and Social Media(ICWSM-2007)*.
- Zhang, E. and Y. Zhang. 2006. "UCSC on TREC 2006 Blog Opinion Mining." *Proceedings of the 15th Text Retrieval Conference(TREC-2006)*.
- Zhang, M. and X. Ye. 2008. "A Generation Model to Unify Topic Relevance and Lexicon-based Sentiment for Opinion Retrieval." *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval(SIGIR-2008)*, 411-418.
- Zhou, G., H. Joshi, and C. Bayrak. 2007. "Topic Categorization for Relevancy and Opinion Detection." *Proceedings of the 16th Text Retrieval Conference(TREC-2007)*.