혼합 방식에 기반한 의견 문서 검색 시스템*

An Opinionated Document Retrieval System based on Hybrid Method

이승욱(Seung-Wook Lee)**
송영인(Young-In Song)***
임해창(Hae-Chang Rim)****

초 록

최근 웹 환경이 대중화되고 개방됨에 따라 웹은 단순한 정보 획득의 공간이 아닌, 의견 표출과 교환의 장이 되어 가고 있으며, 이에 따라 웹 상에서 표출된 특정 주제에 대한 사람들의 의견을 자동으로 검색하기 위한 기술 개발의 필요성이 점차 증대되고 있다. 이러한 의견 문서 검색 문제는 사용자 질의와 문서간의 적합성만을 고려하는 일반적인 정보검색 방법으로는 해결하기 어려우며, 문서 내 의견 포함 여부 분석을 수행할 수 있는 더욱 진보된 시스템을 필요로 한다. 본 논문에서는 기존 검색 시스템의 구조 하에서, 의견 문서 검색을 효과적으로 수행할 수 있는 시스템을 제안한다. 의견 검색을 수행하기 위해 문서 내 의견 분석 방법에 대해 기존의 사전 기반 방식과 기계학습 기반 방식을 결합한 새로운 혼합 방식을 제안하고, 실험을 통하여 검색 성능을 개선하는 효과가 있음을 보였다.

ABSTRACT

Recently, as its growth and popularization, the Web is changed into the place where people express, share and debate their opinions rather than the space of information seeking. Accordingly, the needs for searching opinions expressed in the Web are also increasing. However, it is difficult to meet these needs by using a classical information retrieval system that only concerns the relevance between the user's query and documents. Instead, a more advanced system that captures subjective information through documents is required. The proposed system effectively retrieves opinionated documents by utilizing an existing information retrieval system. This paper proposes a kind of hybrid method which can utilize both a dictionary-based opinion analysis technique and a machine learning based opinion analysis technique. Experimental results show that the proposed method is effective in improving the performance.

키워드: 정보검색, 의견 문서 검색, 혼합 방식
information retrieval, opinion retrieval, hybrid method

---

* 이 연구에 참여한 연구자(의 일부)는 '2단계 BK21사업'의 지원비를 받았음.
** 고려대학교 정보통신대학원 컴퓨터·전파통신공학과(swlee@nlp.korea.ac.kr) (제1저자)
*** 고려대학교 정보통신대학원 컴퓨터학과(song@nlp.korea.ac.kr) (공동저자)
**** 고려대학교 정보통신대학원 컴퓨터·전파통신공학과(rim@nlp.korea.ac.kr) (공동저자)
■ 논문접수일자 : 2008년 11월 13일 ■ 최초심사일자 : 2008년 11월 17일 ■ 게재확정일자 : 2008년 11월 26일
■ 情報管理學會誌, 25(4): 115-129, 2008. [DOI:10.3743/KOSIM.2008.25.4.115]

1. 서 론

최근 수년간 일반 사용자들에 의해 작성한 사용자 제작 콘텐츠(user generated contents)가 크게 각광 받고 있다. 현재 웹 이용자들은 단순한 정보의 소비자에 그치지 않으며, 블로그, 인터넷 게시판, 온라인 뉴스의 댓글 등을 통해 다양한 주제에 대한 자신의 의견을 적극적으로 피력하고 있으며, 또한 이렇게 생성된 의견 정보들은 다른 웹 사용자들의 의사 결정에 있어 참고자료로서 유용하게 활용되기도 한다.

이 의견 정보들은 다른 다수의 이용자들의 의사 결정에 주요한 영향을 미칠 수 있다는 잠재성을 지니고 있다는 점에서 특히 중요하다. 예를 들어, 특정 상품에 대한 사용자들의 평가는 다른 이용자에게 그 상품에 대한 선입관을 가지게 할 수도 있으며, 사용자 평가의 긍·부정 성향에 따라 제품의 판매량의 변화 및 기업의 이윤과 직결되기도 한다. 이에 따라 기업에서는 다수의 소비자의 의견을 빠른 시간 내에 자동으로 파악하여 반영하고자 하는 요구가 증가하고 있다. 또한, 특정 주제에 대한 의견 정보를 검색하고자 하는 사용자 정보 요구 역시 증가하는 추세다.

단순히 문서와 질의의 적합성만을 검색에서 고려하는 현재 일반적인 정보검색시스템으로는 이러한 기업이나 사용자들의 정보 요구를 충족시키는데 한계가 있음은 명백하다. 현행 정보검색시스템은 사용자 질의와 문서 간의 유사도 혹은 적합도에 따라 문서를 순위화한 결과를 제공하는 것을 목표로 하며, 이 과정에서는 의견 정보의 포함 여부는 고려하지 않는다. 따라서 현행 정보검색 방식 하에서는 사용자는 의견 정보를 추려내기 위하여 많은 검색 결과를 일일이 살펴보아야 하는 불편을 피할 수 없다. 따라서 이러한 의견 정보에 대한 사용자 정보검색 요구를 충족시키기 위해서는 의견 정보 검색에 특화된 별도의 검색 방법론 및 시스템의 개발이 필수적으로 요구된다.

본 연구에서는 이러한 의견 정보에 대한 사용자의 정보 요구를 효과적으로 처리할 수 있는 의견 문서 검색 시스템의 개발을 목표로 한다. 본 연구에서 제안하는 검색 시스템은 일반적인 정보검색시스템과는 다르게, 검색 시 사용자의 질의에 적합한 주제의 문서를 검색하되, 의견이 포함된 문서를 우선적으로 제공하도록 설계되었다. 또한, 본 연구에서는 의견의 긍부정 성향에 따라 긍정 의견 문서 검색 결과와 부정 의견 문서 검색 결과를 구분하여 제공할 수 있도록 하는 기법에 대해서도 논한다.

본 논문의 구성은 다음과 같다. 2절에서는 의견 정보 검색에 대한 기존 연구에 대해 기술한다. 다음 3절에서는 제안하는 의견 정보 검색 방법과 이에 기반하는 시스템에 대해 논한다. 이후 4절에서 실험을 통해 제안하는 시스템이 의견 정보 검색에 유용할 수 있음을 보이고, 마지막으로 5절에서는 본 연구의 내용을 요약하고 향후 연구에 대해 논의하는 것으로 결론을 맺는다.

2. 관련 연구

미국 NIST(National Institute of Standards and Technology)의 주관으로 매년 개최되고 있는 검색 관련 유명 학회인 TREC(Text REtriev-

al Conference)은 매년 웹 검색(Web Track), 대용량 질의 검색(Million Query Track), 질의 응답(Question Answering Track), 대용량 정보검색(Terabyte Track) 등과 같은 정보검색에 관련된 다양한 세부 분야를 설정하고, 각 분야에 대해 동일한 평가 집합을 사용하여 다양한 방법론을 실험할 환경을 제공함으로써 정보검색에 대한 심도 있는 논의를 증진시켜 왔다.

이 중 Blog Track은 2006년부터 새로 생긴 분야로, 블로그 문서를 대상으로 하여 특정 주제와 적합하면서 동시에 의견을 포함하고 있는 문서를 찾는 것을 목표로 한다. 매년 많은 연구팀들이 참가하여 다양한 연구 성과를 보고하고 있으며, 각 팀들이 제출한 검색 결과는 수작업으로 평가되어 공개 된다. 이 자료들은 정보검색 및 의견 검색 연구에 중요한 자원으로 활용되고 있다.

의견 정보 검색 분야에 대한 최근의 논의와 연구는 이러한 Blog Track을 중심이 되어 진행되고 있으므로, 본고에서는 이전 Blog Track에 참가하였던 의견 정보 검색 시스템의 방법론을 중심으로 관련 연구에 대해 논하고자 한다.

2.1 2단계 접근법

대다수의 Blog Track 참가 팀들은 질의와 문서와의 적합성을 판별하기 위한 주제별 검색을 수행한 후(주제 검색 단계; topical retrieval phase), 이 후 단계에서 상위로 순위화된 문서들에 대하여 의견 포함 여부를 분석(의견 분석 단계; opinion analysis phase)하는 2단계 접근 방법을 이용하였다(Ounis et al. 2006). 2단계 접근 방법을 사용하는 시스템의 경우 주제별 검색 단계에서는 기존의 검색 모형을 그대로 활용하는 경우가 많았다. 예를 들어, Vechtomova (2007), Java et al.(2006), Joshi et al.(2006), Attardi & Simi(2006)는 정보검색 분야에서 가장 널리 사용되는 검색 모형 중 하나인 BM25 확률 검색 모형을 이용하였으며, Liao et al.(2006), Clark et al.(2006), Yang et al. (2006) 등은 언어 모형(language model) 기반 검색 모형을 활용하였다. Yang et al.(2006)은 역시 일반적인 정보검색 모형 중 하나인 벡터 공간 모형(vector space model)을 이용하였고, Hannah et al.(2007)는 DFR(divergence from randomness) 검색 모형을 사용하여 주제별 검색 단계를 수행하였다. 몇몇 시스템들은 블로그 문서 집합의 특성을 고려하여 별도의 최적화 작업을 추가적으로 수행하기도 하였으나, 기존 적합성 기반 정보검색 방식에서 크게 벗어나지는 않았다.

두 번째 단계인 의견 분석 단계에서는 첫 단계에서 상위 검색된 문서들을 대상으로 하여 의견 포함 여부를 조사, 의견을 포함하는 문서를 상위 순위에 배치되도록 재순위화 된다. 이 단계는 의견 문서 검색에 있어 가장 핵심적인 부분에 해당된다. 이 과정에서 다양한 방법론이 제기되어 왔으며, 크게 다음과 같이 사전 기반 방식과 기계학습 기반 방식으로 구분할 수 있다.

- 사전 기반 방식: 사전 기반 방식은 의견을 기술할 때 자주 사용되는 어휘들을 수집하여 사전으로 구축한 후, 문서 내 사전 등재 어휘들의 출현 정보에 기반하여 문서의 의견 포함 여부를 판단하고 이에 따라 검색 결과를

재순위화하는 방식을 말한다. 이 방식은 많은 기존 연구에서 논의되어 왔으며(Zhou et al. 2007; Vechtomova 2007; Hannah et al. 2007; Liao et al. 2006; Oard et al. 2006), 대부분 사전에 등재된 어휘가 많이 출현할수록, 그리고 질의 단어들과 근접하여 출현할수록 높은 의견 점수(opinion score)를 부여하게 함수를 정의하여 사용한다.

- 기계학습 기반 방식 : 많은 시스템(Zhang & Yu 2006; Attardi 2006; Joshi et al. 2006)에서 문서 분류 모형을 활용하여 의견 문서와 비의견 문서 분류를 수행하였다. 분류기를 학습시키기 위해 자체적으로 의견 정보가 부착된 학습 말뭉치를 구축하였거나, 의견 문서들이 주로 차지할 것으로 예상되는 특정 리뷰 사이트들의 문서들을 자동으로 수집하여 학습 말뭉치로 활용하였다. 분류 모형은 여러 기계학습 기법 중 다양한 분야에서 높은 성능을 보였던 지지벡터모형(support vector model)을 주로 사용하였고(Zhang et al. 2006; Java et al. 2006), 지수형회귀모형(logistic regression model)을 사용하기도 하였다(Zhang & Zhang 2006).

2.2 통합 접근법

두 단계로 나누지 않고 하나의 통합된 모형으로 의견 문서를 검색한 시스템도 제안되었다. Zhang & Ye(2008)는 생성 모형(generative model)관점에서 문서의 의견 포함 여부와 질의에 대한 문서 적합도를 하나의 통합된 모형을 사용하여 검색에 반영하였다. 이 검색 모형은 문서 d가 의견 s를 담고 있고, 질의 q에 대해서 적합할 확률, P(d|q,s)로 문서를 순위화 한다. 이 모형은 질의에 대한 문서의 주제적 적합성(topical relevance)과 문서에서의 의견 생성 확률(opinion generation probability)의 곱으로 검색 결과를 순위화 한다.

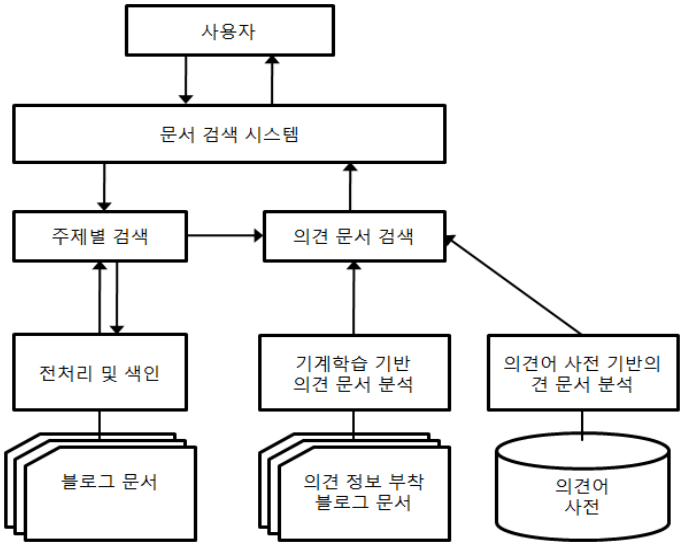
3. 제안 시스템

전술한 바와 같이, 제안하는 시스템은 색인된 블로그 문서로부터 사용자의 질의에 적합한 의견 문서를 검색하여 사용자에게 제공하는 것을 그 목적으로 한다. 〈그림 1〉은 제안하는 시스템의 구조를 개략적으로 보여 준다. 일반적인 검색 시스템과 같이, 본 시스템에서도 블로그 문서는 전처리 과정을 거친 후 색인되고, 주제별 검색 모듈이 사용자의 질의에 적합한 문서를 검색한다. 이 후 질의에 대해 적합하다고 판단되는 상위 검색 문서들을 대상으로 상위로 재순위화 하는 과정을 거친다. 제안하는 시스템의 의견 문서 검색 모듈은 사전 기반 방식과 기계학습 기반 방식을 혼합하여 최종 의견 문서 검색 결과를 산출하는 방식을 사용하며, 이 점이 기존 의견 검색 연구들과의 가장 큰 차이점이라 볼 수 있다. 이후 단락에서는 본 시스템의 주요 구성 요소에 대해 상세 기술한다.

3.1 전처리 및 색인

3.1.1 HTML 태그 제거

제안하는 시스템에서 검색 대상으로 설정한 블로그 문서는 〈그림 2〉에서 볼 수 있듯이



〈그림 1〉 제안하는 시스템의 구조도

```
<table width="100%" border="0" cellspacing="0" cellpadding="1">
<tr nowrap bgcolor="#ccccff">
<td valign="middle">
<A HREF="/cf_lists/index.cfm/method:reply/forumid:4/threadid:43618/`messagecounter:226253"
CLASS="n avbar-left">Reply</A> |
<A HREF="/cf_lists/messages.cfm/forumid:4/threadid:43618" CLASS="navbar-left">Full Thread</A>
</td>
<TD align="right"><b>12/06/2005 02:10 PM</b></TD>
</tr>
<tr>
<td colspan="2">
<b>Author:</b> Michael Dinowitz<BR>
<b>Short Link:</b> http://www.houseoffusion.com/lists.cfm/link=m:4:43618:226253<BR><BR>
I'd be more than happy to accept other peoples sites for a modest fee. :)<br>My cell has the abi
lity to remote desktop into a server which is super when <br>dealing with site down issues.<br
><br><br>&gt;&gt;<br>&gt;&gt; I'm posting my cell phone number for those who have an emergen
cy or if<br>&gt;&gt; someone sees a site problem. Please use it sparingly. Thank you<br>&gt;
<br>&gt; So... um, you mean just your site, right?  ;^)<br>&gt;<br>&gt; Jim Davis<
br>&gt;<br>&gt;<br>&gt;<br>&gt;<br>&gt;
<BR>
</td></tr>
</table>
```

〈그림 2〉 HTML 블로그 문서의 예

HTML(hyper text markup language)로 작성되어 있다. 텍스트 본문이외의 태그 등 질의와 문서간 매칭에 불필요한 부가 요소들은 문서 색인 및 검색에 앞서 전처리 단계에서 제거하였다.

3.1.2 스팸 문서 제거

블로그 문서는 임의의 작성자에 의해 작성되며, 내용 및 구성에 아무런 제한을 받지 않는다. 이에 따라 블로그 문서 집합은 상업적인 목적에 의한 광고성 문서를 다수 포함할 수 있으나 이들 광고성 문서는 질의 대상 정보로서 가치 있는 내용을 포함하는 일이 드물다. 색인 효율성이나 검색 결과에 대한 사용자 만족도를 고려하여 색인 단계 이전에서 광고성 문서를 검출, 제거하는 것이 일반적이며, 본 연구에서도 스팸 필터링 기법을 사용하여 이와 같은 광고성 문서들을 전처리 단계에서 인식, 색인 문서 집합에서 제거하였다.

현재까지 많은 수의 효과적인 스팸 필터링 기법들이 제안되었으나, 제안하는 시스템은 사용 및 구현이 비교적 간단하면서도 높은 정확도를 보인 단순 베이지안 분류 모형(naive Bayesian classification model) 기반 스팸 필터링(spam filtering) 방식을 이용하였다. 이 방식은 단어들의 스팸/비스팸 문서 집합에서의 발생 확률들을 학습 집합으로부터 각각 추정한 후 이 확률들에 기반하여 이후 입력 문서들에 스팸 문서일 확률을 추정하는 방식으로 스팸 문서 분류를 수행한다.

3.2 주제별 검색

본 논문에서는 사용자 질의에 관련된 문서를 해당 질의에 대한 적합도에 따라 정렬하는 것을 목표로 하는 전형적인 정보검색 모형을 주제별 검색 모형으로 지칭한다. 주제별 검색 모형으로는 벡터 공간 모형(vector space model), 확률 기반 검색 모형(probabilistic retrieval model) 등의 다양한 방법론들이 제안되었다. 본 연구에서는 주제별 검색을 위하여 공개된 정보검색시스템 중 하나인 Lemur[1]를 사용하였으며, 검색 모형으로는 최근 널리 사용되고 있고, 또한 성능이 우수하다고 알려진 검색 모형 중 하나인 언어 모형 기반 검색 모형(language model for IR)을 사용하였다(Ponte & Croft 1998).

3.3 질의 구성

본 연구에서 제안하는 검색 시스템에서는 주제별 검색의 성능을 보다 향상시키기 위해 원 사용자 질의에 대해 이와 연관된 새로운 단어들을 추가하는 작업을 질의 처리 단계에서 수행한다.

먼저, 〈그림 3〉은 본 시스템에서 입력으로 사용하는 Blog Track 형식의 질의 예를 보여준다. 그림에서 볼 수 있듯이, 입력 질의는 서로 다른 3가지 필드(Title, Description, Narrative)로 구성되어 있다. Title은 일반적으로 한 두 단어들로 짧게 구성되어 있지만 질의의 주제와 가장 밀접한 연관이 있다. 다른 두 Description과 Narrative 필드는 Title 필드의 내용에 대한 보충 기술을 담는다. 여러 이전 연구들에서 Narrative 정보는 유용하지 않은 반면 Description 정보를 활용하였을 때는 성능 향상을 이루었다고 보고하였기 때문에, 본 시스템에서는 Title 정보와 Description 정보를 사용

1) 〈http://www.lemurproject.org〉

<top>
<num> Number: 884

<title> Qualcomm

<desc> Description:
What are opinions of Qualcomm, its business practices, marketing and products?

<narr> Narrative:
Any opinion of Qualcomm's business practices, marketing procedures or products are relevant. Statements of mergers or takeovers are relevant only if they serve to alter the opinion of the company. Any reference to Qualcomm Stadium used by the San Diego Chargers is not relevant.

</top>

〈그림 3〉 입력 질의 예

자의 입력 질의로 간주하였다. 또한 검색 시 사용자 질의에서 제거리스트(stopword list)에 해당되는 단어들은 제외하였다.

3.4 의견 문서 검색

이 절에서는 의견 문서 검색 시스템을 구축하기 위해 적용한 의견 문서 분석 방법 및 이에 기반한 재순위화 기법에 대해 기술한다. 제안하는 시스템에서의 의견 문서 분석은 기존 사전 기반 의견 정보 분석 방법과 기계학습 기반 방법의 분석 결과를 결합하여 사용하는 혼합형 방식을 취했다는 점에서 기존 관련 연구에서 사용한 방법과 크게 구분된다.

3.4.1 사전 기반 의견 정보 분석

(1) 사전 구축

사전 기반 방식은 의견을 표현할 때 사용되는 단어 어휘들을 수집한 후, 이 단어들의 출현에 따라 의견의 포함 여부를 판단한다. 따라서 사전 기반 방식의 성능 향상을 위해서는 적용률이 높은 양질의 의견어 사전 구축이 필수적 요소 중 하나이다. 이에 본 연구에서는 이전 사전 기반 의견 분석 연구에서 사용된 복수의 의견어 사전들을 통합하여 확장된 의견어 사전을 구축 및 활용하였다. 다음 세부 항목에서는 시스템에서 최종 의견어 사전 구축에 사용된 개별 의견어 사전 목록 및 구축방법에 대해 간략히 기술한다.

- General Inquirer : 수작업으로 수집된 의견어 어휘 사전인 General Inquirer는 (Stone & Dunphy 1966)에 의해 수동 구축되었으며, 다양한 의미적 분류와 감정 분류로 구성되어 있다. 이 중 의견과 감정 분류에 포함된 단어만을 선택하여 의견어 어휘 사전으로 사용하였다.

- WordNet : WordNet(Miller 1992)은 대표

〈표 1〉 의견 사전별 어휘수

의견 사전	어휘수
GI	3,621
WordNet	16,699
Wilson	6,777
통합	24,172

적인 범용 온라인 사전 중 하나로 많은 자연어처리에 활용되고 있다. 본 연구에서는 범용 사전인 WordNet으로부터 의견어 사전을 구축하기 위하여, Kim & Hovy(2004)에서 제안한 WordNet 기반 의견어 사전확장 방법을 사용하였다. 이 방법에서는 소량의 의견어 집합을 수동으로 구축한 후, WordNet의 동의어, 상하위어 및 반의어 관계를 사용하여 의견어 사전을 점진적으로 확장한다. 이 작업은 더 이상 확장 가능한 단어가 없을 때까지 반복된다.

- Wilson Word Set : 의견 문장을 탐지하기 위한 주관성을 나타내는 표현을 수집한 사전 Wilson et al.(2003)이며, 변형 없이 그대로 개별 단위 의견어 사전으로 사용하였다.

최종 통합된 의견어 사전은 이들 세 의견어 사전에 등재된 의견어 어휘의 합집합으로 구성된다. 이 과정에서 중복 어휘는 제거된다. 각 사전의 어휘수는 〈표 1〉에 나타난 바와 같다.

(2) 사전 등재 어휘 가중치 계산

의견어 사전 등재 어휘는 의견 문서 분석에 있어 서로 다른 중요도를 지닐 수 있다. 예를 들어, 'may'와 같은 단어는 의견을 표현하기 위해서도 사용되지만, 다른 용도로도 빈번히 사용된다. 그러나 'excellent'와 같은 단어는 주로 의견을 표현하기 위해서만 사용되는 경향이 있다. 따라서 이처럼 서로 다른 특성을 고려하기 위하여 의견어 사전의 각 어휘에 의견 분석 시 중요도에 따라 서로 다른 가중치를 부여하였다.

각 단어에 대한 가중치는 의견을 포함한 문서에 자주 출현하면서도 의견을 포함하지 않은 문서에서는 출현하지 않는 어휘들에 대해 높은 가중치가 부여 되도록 계산하였다. 2006년, 2007년에 공개된 의견 포함 여부의 정보가 표기되어있는 블로그 문서를 가중치 부여를 위해 사용하였으며, 가중치 계산 수식은 Dave et al.(2003)가 제안한 방법을 사용하였다. 구체적인 가중치 부여 수식은 다음 〈수식 1〉과 같다.

$$weight(w_i) = \frac{P(w_i|O) - P(w_i|N)}{P(w_i|O) + P(w_i|N)} \qquad (1)$$

w_i는 의견어 사전에서 i번째 항목을 나타낸다. O는 의견을 포함하고 있는 문서를, N은 의견을 포함하고 있지 않는 문서를 의미한다. 의견 문서에 어휘가 출현할 확률보다 비의견 문서에서 출현할 확률이 더 높을 때는 가중치가 음수 값인 특징이 있다.

(3) 사전 기반 의견 정보 분석

가중치가 계산된 의견어 사전은 주제별 검색엔진의 검색 결과를 대상으로 의견 정보를 분석하여 재순위화 하는데 사용된다.

일반적으로 사전 기반 의견 정보 분석 방식에서는 의견어 사전에 등재된 어휘가 문서에 출현한 정도에 따라 해당 문서가 의견을 포함하고 있을 가능성을 나타내는 의견 점수(opinion score)

를 계산하고, 이 의견 점수를 사용하여 주제별 검색 결과를 재순위화 한다.

본 연구에서 사용한 의견어 출현에 따른 문서의 의견 점수 계산 방법은 다음 〈수식 2〉와 같다.

$$score_{lex}(d) = \sum_{w_i \in d} \frac{tfidf(w_i) \times weight(w_i)}{shortest(w_i, Q)} \quad (2)$$

수식 (2)는 기본적으로 가중치가 높은 의견어가 많이 출현한 문서일수록, 또 문서에서 출현한 의견어와 사용자 질의어 사이의 거리가 짧을수록 높은 의견 점수가 부여되도록 설계되었다. 수식에서 *tfidf*는 현재 문서 내 어휘 w_i의 출현 빈도와 전체 문서 집합에서 w_i가 출현한 문서의 수의 역수를 곱하여 계산된다. *shortest* 함수는 질의 Q를 구성하고 있는 각 질의어 q중 w_i와 가장 가까운 단어 q'를 찾아 그 사이의 거리를 반환 한다. 거리는 단어의 수로 정의된다. 이때 정확한 점수 부여를 위해 질의는 Title 필드의 단어만으로 한정하였다.

3.4.2 기계학습 기반 의견 정보 분석

최근 기계학습 기법, 특히 자동 분류기법은 언어 처리의 다양한 분야에 적용되어 효과적으로 활용되고 있다. 일반적으로 기계학습 기법은 분류 대상이 되는 표지(label)가 부착된 예제들로 구성된 학습 집합을 구축한 후, 이 데이터를 이용하여 분류 모형을 자동으로 학습한다. 이 후 표지가 부착되지 않은 새로운 예제가 주어졌을 때 이의 표지를 예측한다. 최대 엔트로피 모형(maximum entropy model), 신경망(neural network), 단순 베이지안 분류모형(naive Bayesian model) 등 다양한 종류의 기계학습 기법이 제안되었으며, 많은 분류 문제에 효과적이라고 알려져 있는 지지 벡터 기계(support vector machine; SVM)는 의견 문서 분류에도 높은 성능을 보여 왔다. 본 연구에서도 의견 문서 분류기로 SVM을 이용하였으며[2] 문서 내 출현한 의견어 사전의 등재 어휘를 자질로 사용하였다. SVM 이진 분류기의 결과는 실수 값으로, 양수이면 의견 문서로 분류되고 음수이면 비의견 문서로 분류된다. 〈수식 3〉과 같이 문서의 재순위화를 위하여 표지에 대한 SVM의 출력 신뢰도 값을 해당 문서에 대한 의견 점수로 간주하여 사용하였다.

$$score_{2-svm}(d) = SVM(d) \quad (3)$$

3.4.3 사전 기반 분석 결과와 기계학습 기반 분석 결과의 혼합

의견어 사전 기반 분석 방식과 기계학습 기반 분석 방식은 주제별 검색 결과를 대상으로 의견 포함 여부를 분석하기 위해 활용된다. 〈수식 4〉와 같이 주제별 적합성, 의견어 사전 기반 점수, 기계학습 기반 점수들을 결합하여 의견 점수를 계산한다. 기존의 시스템들은 사전 기반 방식이나 기계학습 기반 방식 중 하나를 택하였지만 제안하는 시스템에서는 모두 고려하게 설계하였다. 수식에 따르면 주제에 관련 깊은 문서라도 의견을 포함하고 있지 않다면 상대적으로 낮은 점수가 부여된다.

2) 본 연구에서는 실험을 위하여 공개된 지지 벡터 학습기 툴킷인 svmlight (http://svmlight.joachims.org)을 사용하였다.

$$score_{op}(d) = \lambda_1 score_{init}(d) + \lambda_2 score_{lex}(d) + \lambda_3 score_{2-svm}(d) \quad (4)$$

단, $\lambda_1 + \lambda_2 + \lambda_3 = 1$

수식에서, $score_{init}$는 주제별 검색에서 문서 d에 할당한 적합성 점수를 의미하며, $score_{lex}$는 수식 (2)를 사용한 사전 기반 분석 방식의 의견 점수를 의미한다. $score_{2-svm}$은 기계학습 기법에 기반한 수식 (3)의 의견 점수를 의미한다. 이 점수들을 결합하는데 사용된 가중치 상수 λ의 값들은 반복 실험을 통해 경험적으로 결정하였다.

의견 포함의 여부가 고려된 검색 결과를 생성하기 위해 의견 점수가 높은 순으로 정렬하는 과정을 거친다. 이때 의견 점수가 미리 정의된 임계값보다 작을 때는 검색 결과에서 제거하여 보다 정확한 결과를 도출한다.

3.5 긍·부정 의견문서 검색

경우에 따라 사용자는 단순히 질의의 주제에 대한 의견을 담고 있는 문서를 찾는 것에서 벗어나, 주제에 대한 긍정 의견, 부정 의견 등 특정 부류의 의견을 담고 있는 문서들을 찾고자 할 수도 있다. 예를 들어, 특정 상품이나 대통령 선거의 입후보자에 대한 긍정 의견이나 부정 의견에 대한 정보만을 궁금해 하거나 원하는 사용자도 있을 수 있을 것이다. 이러한 측면에서 긍부정 의견 검색은 의견 검색의 하위 문제로서 중요히 취급되어 왔고, 앞서 언급한 Blog Track에서도 하나의 독립된 세부분야로 연구가 진행되고 있다.

〈표 2〉 Blogs06 문서 집합의 통계

	수	용량
Feed	100,649	38.6GB
Permalink	3,215,171	88.8GB
Homepage	324,880	20.8GB

본 연구의 제안 시스템에서도 의견의 긍부정에 따라 서로 다른 두 검색 결과를 별도로 구성하는 작업을 수행하였다. 3.4절에서 기술한 의견 검색 방식과 유사하게 사전 기반 점수와 기계학습 기반의 점수를 결합하여 긍부정 의견 문서 재순위화에 사용한다. 사전 기반 의견 점수는 3.4절에서 기술된 것과 동일하나, 긍정/부정 의견 문서를 구분하기 위해 의견/비의견 문서 부류를 결정하는 이진 SVM 분류 모형이 아닌 SVM을 이용한 삼항 분류 모형의 결과를 기계학습 기반 의견 점수로 사용하였다. 이 분류 모형은 문서를 긍정, 부정, 혼합 등의 세 가지로 분류한다. 혼합 부류는 문서 내 긍정 의견과 부정 의견이 혼재된 상태를 의미하며, 최종 검색 결과에서는 제거된다.

긍정 혹은 부정 의견 문서 검색을 위한 재순위를 위해 다음과 같이 순위화 수식을 정의하여 사용하였다.

$$score_{pol}(d) = \lambda_4 score_{op}(d) + \lambda_5 score_{3-svm}(d) \quad (5)$$

단, $\lambda_4 + \lambda_5 = 1$

특정 문서에 대한 〈수식 5〉의 점수가 양수이면 긍정 의견, 음수이면 부정 의견 문서임을 의미한다. 이에 기반하여 긍정 의견에 대한 문서 검색 결과와 부정 의견에 대한 문서 검색 결과를 별도로 산출하여, 성능을 평가하였다.

4. 실험 및 평가

4.1 실험 데이터

4.1.1 문서 집합

제안하는 시스템에서 실험에 사용한 문서 집합으로 TREC(Text REtrieval Conference)의 Blog Track에서 이용되고 있는 블로그 문서 집합(Blogs06)을 사용하였다. 이 문서 집합은 University of Glasgow에서 2005년 말에서 2006년 초까지 자동으로 수집되었다(Macdonald et al. 2006). 수집된 블로그는 대다수가 영어로 작성되었고 그 외의 다양한 언어들을 포함하고 있다. 〈표 2〉는 블로그 문서 집합에 대한 통계이다. 또한 이 문서 집합은 약 100,000개 정도의 광고성 문서 혹은 스팸 문서를 포함하고 있다고 추정된다.

4.1.2 질의 집합

실험을 위해 Blog Track에서 2007년에 사용한 50개의 질의를 사용하였다. 이 질의들은 *"steve jobs", "davos", "allianz", "winter olympics"*등과 같이 인명, 지명, 조직명, 제품명, 이벤트 등에 관한 것이며, 〈그림 3〉에 나타난 예제와 같이 Title, Description, Narrative 등 3 개의 파트로 구성되어 있다.

4.1.3 적합 문서 집합

본 시스템의 성능을 평가하기 위한 적합 문서 집합으로 2006년, 2007년에 Blog Track에 참가한 여러 팀들이 제출한 결과를 바탕으로 TREC에서 수작업으로 평가하여 공개한 평가 집합을 사용하였다. 이 평가 집합은 각 참가팀이 제출한 질의 당 1,000개의 의견 문서 검색 결과에 대해 평가자들이 질의 주제에 대한 적합성 여부, 질의 주제에 대한 의견 포함 여부, 긍부정 의견 여부 등을 수동으로 판단하여 구축된 집합이다. 〈표 3〉은 적합 문서 평가 집합의 통계를 보여준다.

4.2 평가 척도

제안한 의견 정보 검색 시스템의 성능 평가을 위해 정보검색 분야에서 주요 평가 척도로 사용되고 있는 Mean Average Precision(MAP), R-Precision(R-Prec), Precision at 10(P@10) 등을 사용하였다.

〈표 3〉 적합 문서 평가 집합의 통계

구 분	레이블	수	비 율
평가 불가	-1	0	0%
비적합	0	89,925	73%
적합	1	13,548	11%
긍정 의견 포함	2	5,551	4%
긍정 부정 의견 모두 포함	3	5,860	4%
부정 의견 포함	4	7,119	5%
합	-	12,2003	100%

4.3 실험 및 평가

본 절에서는 제안 시스템에 대한 실험 및 그에 대한 성능 평가에 대해 기술한다. 본 논문에서는 주제별 검색 결과, 의견 검색 결과, 그리고 긍부정 의견 검색 결과에 대해 각각에 대해 개별적으로 성능을 평가한 후, 그 결과를 분석하였다.

4.3.1 의견 검색 성능 평가

〈표 4〉와 〈표 5〉는 각각 Title 질의와 Title+Description 질의에 대해, 제안하는 의견 검색을 위한 재순위화 방법을 적용한 실험 결과를 보여준다. 이들 표에서 나타난 것처럼 사전 기반 분석 기법만을 사용했을 경우($\lambda_2=0.7$, $\lambda_3=0.0$)나 기계학습 기법만을 사용했을 경우($\lambda_2=0.0$, $\lambda_3=0.7$)보다, 사전 기반 의견 분석 모형과 기계학습 기반 의견 분석 모형을 모두 사용했을 때($\lambda_2=0.1$, $\lambda_3=0.6$) 가장 높은 성능을 보였다.

또한, 〈표 4〉와 〈표 5〉의 비교에서 알 수 있듯이 주제별 검색 성능의 개선은 최종 의견 검색 성능 개선에 있어서도 중요한 것으로 관찰되었다. 재순위화를 하지 않았을 경우, Title 질의를 사용한 경우와 주제별 검색 성능 비교에서 그보다 좋은 성능을 보였던 Title+Description 질의를 사용한 경우 사이의 성능 차이는 미미하였지만, 의견 문서 검색을 위한 재순위화를 수행하였을 경우, 약 10% 정도의 성능 차이가 있었다.

〈표 6〉은 의견 문서 검색을 수행한 후, 의견의 긍부정 정보를 추가적으로 수행한 성능 변화를 보여준다. 긍부정 의견 검색의 성능은 〈표 4〉와 〈표 5〉에서의 의견 문서 검색의 성능에 비해 상당히 하락하였음이 관찰되었는데, 이는

〈표 4〉 Title 정보만을 질의 구성에 이용하였을 때의 의견 문서 재순위화 성능

의견 문서 재순위화 모형	MAP	R-Prec	P@10
재순위화 안함	0.2112	0.2922	0.3740
$\lambda_1=0.3$, $\lambda_2=0.7$, $\lambda_3=0.0$	0.2138	0.2945	0.3900
$\lambda_1=0.3$, $\lambda_2=0.0$, $\lambda_3=0.7$	0.2729	0.3399	0.5880
$\lambda_1=0.3$, $\lambda_2=0.1$, $\lambda_3=0.6$	**0.2776**	**0.3497**	**0.6040**

〈표 5〉 Title 정보와 Description 정보를 질의 구성에 이용하였을 때의 의견 문서 재순위화 성능

의견 문서 재순위화 모형	MAP	R-Prec	P@10
재순위화 안함	0.2171	0.2915	0.4140
$\lambda_1=0.3$, $\lambda_2=0.7$, $\lambda_3=0.0$	0.2183	0.2889	0.4060
$\lambda_1=0.3$, $\lambda_2=0.0$, $\lambda_3=0.7$	0.2735	0.3444	0.5940
$\lambda_1=0.3$, $\lambda_2=0.1$, $\lambda_3=0.6$	**0.2763**	**0.3469**	**0.6060**

〈표 6〉 긍부정 의견 검색 성능($\lambda_4=0.8$, $\lambda_5=0.2$)

질 의	MAP	R-Prec	P@10
Topic	**0.1879**	0.3990	0.3596
Topic+Description	0.1775	**0.4010**	**0.3638**

〈표 7〉 제안하는 의견 문서 검색 시스템과 기존 시스템의 성능 차이

모 형	재순위화 안함	재순위화 함	상승률
UGlasgow	0.2817	**0.3264**	15.87%
IndianaU	0.2537	0.2894	14.07%
UArkansas	0.2554	0.2911	13.98%
DalianU	**0.2890**	0.3190	10.38%
UWaterloo	0.2486	0.2631	5.83%
제안하는 시스템	0.2112	0.2776	**31.44%**

긍·부정 의견 검색에서 추가적으로 요구되는 긍부정 의견 문서 분류의 오류가 이후 의견 문서 재순위화 과정에 전파된 것이 주요 원인으로 보인다.

〈표 7〉에 나타난 바와 같이 제안하는 의견 문석 검색 시스템은 기존의 시스템들에 비해 훨씬 높은 성능 상승률을 보였으며, 이는 제안하는 재순위화 방법론이 효과적임을 나타낸다. 제안하는 시스템의 검색 성능은 기존 시스템의검색 성능보다 낮은 성능을 보였고, 이로 인해 최종 검색 성능도 감소하였다. 하지만 본 연구에서는 주제별 정보검색 단계를 연구의 주목적으로 정하지 않았기 때문에 당연한 결과이며, 보다 향상된 정보검색 기법들을 활용한다면 보다 높은 최종성능에 도달할 것으로 예상된다.

5. 결 론

최근 의견 문서 검색의 필요성이 대두됨에 따라, 이를 위한 다양한 방법론들이 연구, 개발되어 왔다. 본 연구에서는 의견 문서 검색을 위해, 2 단계 검색 구조에 기반하여 의견 문서 검색 시스템을 개발하였다. 특히, 의견 문서 분석 및 재순위화를 위하여 기존 사전 기반 방식과 기계학습 기반 방식을 혼합하여 사용하는 방법을 제안하였으며 실험을 통해 제안하는 방법이 의견 검색에 효과적일 수 있음을 보였다.

추후 연구로서, 한국어 문서를 대상으로 의견 검색 시스템을 구축하는 것을 고려해 볼 수 있다. 이를 위해 한국어 의견어 사전의 효과적 구축 방안과 한국어에 적합한 의견 문서 분석 방법 등의 연구가 필요할 것으로 사려된다.

참 고 문 헌

Attardi, G., and M. Simi. 2006. "Blog Mining through Opinionated Words." *Proceedings of the 15th TREC*

Clark, M., U. C. Beresi, S. Watt, and D. Harper. 2006. "RGU at the TREC Blog Track." *Proceedings of the 15th TREC*.

Dave, K., S. Lawrence, and D. M. Pennock. 2003. "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification for Product Reviews."

WWW.

Hannah, D., C. Macdonald, J. Peng, B. He, and I. Ounis. 2007. "University of Glasgow at TREC 2007: Experiments in Blog and Enterprise Tracks with Terrier." *Proceedings of the 16th TREC*.

Java, A., P. Kolari, T. Finin, A. Joshi, and J. Martineau. 2006. "The BlogVox Opinion Retrieval System." *Proceedings of the 15th TREC*

Joshi, H., C. Bayrak, and X. Xu. 2006. "UALR at TREC: Blog Track." *Proceedings of the 15th TREC*.

Kim, S. M., and E. Hovy. 2004. "Determining the Sentiment of Opinions." *Proceedings of Conference on Computational Linguistics (COLLING-04)*

Liao, X., D. Cao, S. Tan, Y. Liu, G. Ding, and X. Cheng. 2006. "Combining Language Model with Sentiment Analysis for Opinion Retrieval of Blog-Post." *Proceedings of the 15th TREC*

Macdonald, C. and I. Ounis. 2006. "The TREC Blog06 Collection : Creating and Analysing a Blog Test Collection." *DCS Technical Report TR-2006-224*. Department of Computing Science, University of Glasgow.

Miller, G. A. 1992. "WordNet: A Lexical Database for English." *Proceedings of the workshop on Speech and Natural Language, ACL*

Oard, D., T. Elsayed, J. Wang, and Y. Wu. 2006. "TREC-2006 at Maryland: Blog, Enterprise, Legal and QA Tracks." *Proceedings of the 15th TREC*.

Ounis, I., M. D. Rijke, C. Macdonald G. Mishne, and I. Soboroff. 2006. "Overview of the TREC-2006 Blog Track." *Proceedings of the 15th TREC*pp.17-31.

Ponte, J. M. and W. B. Croft. 1998. "A Language Modeling Approach to Information Retrieval." *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (SIGIR '98)*

Stone, P. J. and D. C. Dunphy. 1966. *A Computer Approach to Content Analysis*. MIT Press.

Vechtomova, O. 2007. "Using Subjective Adjectives in Opinion Retrieval from Blogs." *Proceedings of the 16th TREC*.

Winson, T., D. R. Pierce, and J. Wiebe. 2003. "Identifying Opinionated Sentences." ACL.

Yang, H., L. Si, and J. Callan. 2006. "Knowledge Transfer and Opinion Detection in the TREC2006 Blog Track." *Proceedings of the 15th TREC*

Yang, K., N. Yu, A. Valerio, and H. Zhang. 2006. "WIDIT in TREC-2006 Blog track." *Proceedings of the 15th TREC*.

Zhang, E. and Y. Zhang. 2006. "UCSC on TREC 2006 Blog Opinion Mining."

Proceedings of the 15th TREC.

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*

Zhang, W. and C. Yu. 2006. "UIC at TREC 2006 Blog Track." *Proceedings of the 15th TREC*

Zhou, G. X., H. Joshi, and C. Bayrak. 2007. "Topic Categorization for Relevancy and Opinion Detection." *Proceedings of the 16th TREC*