TRIB: 웹블로그 댓글분류 시각화 시스템

배민정*, 이윤정*, 지정훈*, 우균*, 조환규*
*부산대학교 컴퓨터공학과
e-mail:{mjbae, leeyj01, jhji, woogyun, hgcho}@pusan.ac.kr

TRIB: A Clustering and Visualization System for Responding comments on WebBlog

Min-Jung Bae*, Yun-Jung Lee*, Jeong-Hoon Ji*, Gyun Woo*, Hwan-Gyu Cho*
*Dept of Computer Engineering, Pusan National University

요 약

최근 들어 인터넷 게시판이나 개인 블로그 등은 온라인상에서 사람들의 정보 공유나 의견 교환의 중요한 매체가 되고 있다. 많은 수의 블로그들은 현재 사회적으로 이슈가 되는 여러 문제들을 반영하고 있다. 또한 최근 댓글을 통해 적극적으로 자신의 의사 표현하거나 다른 사람들의 의견을 살피는 인터넷 사용자의 증가로 인터넷 뉴스나 블로그 기사에 많은 수의 댓글이 달리고 있다. 그러나 대부분의 블로그나 인터넷 포털 사이트의 경우 기사나 댓글들을 순차적인 목록 형태로 제공하므로 자신이 원하는 내용의 댓글을 검색하거나 전체 댓글에 대한 전반적인 파악은 힘든 일이다. 따라서 본 논문에서는 기사에 달린 많은 수의 댓글들을 분류하고, 이를 시각화 하는 시스템인 TRIB(Telescope for Responding comments for Internet Blog)을 제안한다. TRIB은 미리 정의된 사용자 정의 사전을 이용하여 댓글을 내용에 따라 분류하여 시각화 하므로 사용자들은 자신의 관심과 흥미에 따라 개인화 된 뷰를 볼 수 있다. 1,000개 이상의 댓글을 가진 뉴스 기사들을 대상으로 한 실험을 통해 TRIB 시스템의 댓글 분류와 시각화 성능을 보인다.

1. 서론

현재 인터넷 사용자들은 인터넷 포탈이나 블로그 등에서 제공하는 기사를 읽고 정보를 얻는 것뿐만 아니라 댓글을 통해 타인의 의견을 살피거나 자신의 생각을 좀 더 적극적으로 나타내고 있다. 댓글은 누군가가 인터넷에 올린 원문에 대하여 짧게 답하여 올리는 글로 reply, comment와 같은 용어로 사용된다. 2006년 한국인터넷진흥원의 조사에 따르면 조사대상자의 84.8%가 각종 게시물에 달린 댓글을 읽고 있는 것으로 나타났으며, 댓글 이용자 중 절반 이상이 자신의 생각을 표현하거나 타인의 의견을 알기 위해서 댓글을 이용하는 것으로 조사되어, 댓글이 인터넷 이용자들의 생각이나 의견 표현 및 공유 수단임을 알 수 있다[1]. 또한 인터넷 뉴스의 기사에서 댓글은 이용자에게 일종의 신호(signal) 역할을 할 수 있다. 뉴스 기사에서 댓글이 있음으로 해서 다른 사람들도 그 뉴스에 주목하였다는 점을 가시적으로 알 수 있게 해 준다. 사람들이 댓글을 보던 보지 않던, 댓글 자체가 존재함으로써 기사는 좀 더 주목할 만하고, 중요한 것으로 인식될 수 있다[2]. 인터넷 게시물에 대한 댓글들의 수는 기사의 중요도나 관심 정도에 따라 다르긴 하지만 적게는 수백 개에서 많게는 몇 만개 이상이 되기도 한다.

최근 댓글 이용자가 늘어남에 따라 기사 내용과 관련 없는 광고성 글이나 비속어 등이 사용된 악성 댓글들도 다수 포함되어 있어 사회 문제가 되기도 한다. 기존의 포털 사이트나 블로그에서는 이러한 댓글들을 리스트 형태로 보여주기 때문에 많은 수의 댓글이 달린 경우에는 원하는 내용이 포함된 댓글 검색하거나 댓글들의 연관성을 파악하는 것은 상대적으로 어려운 일이다. 효율적인 의견 교환이나 공유를 위해서는 많은 수의 댓글들을 사용자가 원하는 기준으로 필터링하고 이를 시각화하는 도구가 필요할 것이다.

본 논문에서는 인터넷 뉴스나 블로그 기사에 달린 많은 수의 댓글들을 사용자 정의 사전을 통해 내용에 따라 분류하고 이를 시각화하는 시스템인 TRIB을 제안한다. TRIB에서는 화면 중심에 기사를 배치하고 기사의 내용과 연관 정도에 따라 사용자 정의 사전의 단어들을 그 주변에 배치한다. 댓글들은 자신과 가장 의미적으로 연관도가 높은 단어에 속하게 되고 그 단어의 주변에 배치된다. TRIB의 화면 구성은 Nguyen[3]의 연구에서와 마찬가지로 태양계와 유사하다. TRIB은 댓글의 내용에 따른 분류뿐만 아니라 작성된 시간을 기준으로 한 순차적 접근도 가능하므로 논쟁의 경우와 같이 주고받는 형태의 경우 효율적인 검색이 가능하다.

본 논문의 구성은 다음과 같다. 2장에서는 블로그나 웹 검색 결과의 시각화에 대한 관련 연구를 살펴본다. 3장에서는 제안 시스템의 댓글 분류와 시각화 방법에 대해서 자세히 설명한다. 4장에서는 실험 결과를 보이고, 마지막으로 5장에서 결론을 맺는다.

2. 관련 연구

Harris[4] 등은 블로그 시각화 방법인 "We feel fine"이라는 시스템을 개발하였다. 일정 시간마다 전 세계에서 게시되는 블로그 기사들을 수집하여 Madness, Murmurs, Montage, Mobs, Metrics, Mounds와 같이 6가지의 표현방법으로 구분하여 표현한다. 이 시스템은 많은 기사들을 표현하고 있으나 어떤 블로그에서 포스팅 되었는지 혹은 기사들의 앞, 뒤 연결을 알 수 없다. 이것과 유사하게 BBC에서는 Spectrum이라는 news에 대한 댓글을 시각화하는 방법을 개발하였다[5]. BBC 2's White 시즌 중 토론을 조사하여 감정, 지역, 성별 등에 따라 댓글을 클러스터링하고 이를 시각화한다. Spectrum은 감정, 지역, 성별 등과 같이 그룹화 할 기준을 선택할 수 있는 사용자 인터페이스를 제공하고 있어 원하는 기준으로 댓글들을 필터링할 수 있고 움직이는 입자를 클릭하면 토론에서 사용된 댓글을 볼 수 있다. 그러나 "We Feel Fine" 시스템과 마찬가지로 Spectrum도 한 블로그나 뉴스 카테고리 내에서 한 항목에 대한 앞, 뒤 순서나 연결 상태를 알 수 없으며, 실제 관심 이슈에 대한 내용을 담고 있는 글을 찾기 힘들다. Indratmo[6] 등은 Blog 시각화 도구인 iBlogVis 시스템을 제안하였다. iBlogVis에서는 블로그 기사들을 포스팅 시간, 댓글 수, 기사 분량 등에 따라 시각화하고 있으나, 시각화 된 화면에서 실제 기사의 내용을 알 수 없으며 단지 블로그에 대한 개요만 제공한다.

<표 1> 기존 시스템 비교

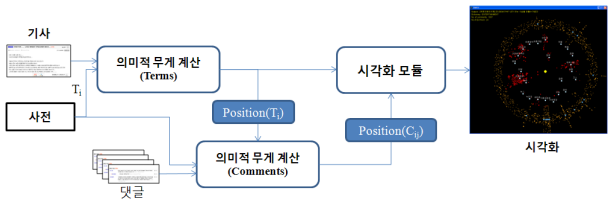
시스템명	응용분야	계층화	클러스터
We Feel Fine[4]	Blog	N	Y
Spectrum[5]	Comments	N	Y
iBlogVis[6]	Blog	N	Y
Takama's[7]	Blog	Y	N
Gregory's[8]	Blog	N	N
TRIB	Comments	Y	Y

표 1은 웹과 관련된 기존의 시각화 시스템들 중에서 블로그와 댓글을 다룬 시스템들의 기능을 비교한 것이다. 웹 검색 결과나 블로그에 대한 시각화에 대해서는 많은 연구가 있었으나 댓글에 대한 검색이나 시각화에 대한 연구는 찾아보기 어려웠다.

3. TRIB의 개요

본 논문에서는 온라인 뉴스 기사에 달린 많은 양의 댓글을 사용자 정의 사전을 이용하여 내용에 따라 분류하고 이를 시각화 하는 시스템인 TRIB을 제안한다. TRIB의 구성은 그림 1과 같다.

사용자 정의 사전은 관심 분야의 단어들을 주제별로 모아서 만들게 되며, 사전의 단어들은 댓글 분류를 위한 키워드로 사용된다. 사용자의 관심 주제별로 정치, 경제, 연예 등 여러 개의 사전을 생성할 수 있다. 댓글은 사전에 정의된 키워드와의 의미적 연관 정도에 따라 분류된다. TRIB의 시각화에서 화면 중심에 키워드가 배치되고 키워드 주변으로 그 키워드에 속한 댓글들이 위치한다.

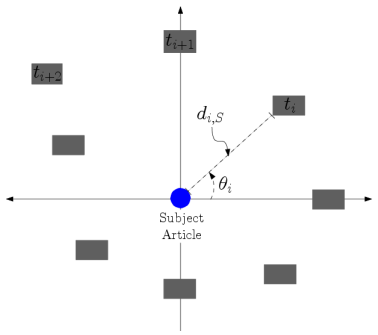


(그림 1) 시스템 구조

키워드는 기사와의 의미적 연관 정도에 따라 연관도가 높을수록 화면 중심에 가깝게 배치된다. 단어 집합 T에 속하는 단어 t_i와 문장 S와의 의미적 연관도를 계산하기 위하여 본 논문에서는 식(1)과 같이 정의하였다.

$$w(t_i,S) = f_i / \sum_{k=1}^{n} f_k, \quad \forall_{t_i} \in T$$
$$0 \le w(t_i,S) \le 1 \qquad (1)$$

여기서, f_i는 t_i가 S에서 나타난 횟수이며, n은 T에 속하는 단어의 개수이다. 사전에 정의된 단어들과 기사내용과의 의미적 연관도가 구해지면 그림 2와 같이 기사를 중심으로 단어들이 방사형으로 배치된다.



(그림 2) 사전에 정의된 단어들의 배치

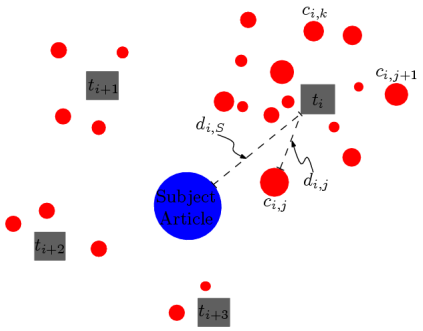
그림에서 θ_i와 $d_{i,S}$는 식 (2)와 같이 계산된다. 여기서, R은 t_i가 위치될 수 있는 최대 반지름이며, N은 사전에 정의된 총 단어 수, 그리고 상수 c는 중심으로의 밀집도를 조절하는 인수이다. 그림 2에서 화면 중심에 가까울수록 기사에 자주 나타나는 단어임을 의미한다.

$$d_{i,S} = R \cdot \exp(-c \cdot w_i)$$
$$\theta_i = 2\pi \cdot i / N \qquad (2)$$

단어들의 위치가 결정되면 댓글들은 자신과 가장 의미적으로 가까운 단어들 주변에 배치된다. 각각의 댓글과 가장 의미적으로 가까운 단어를 찾기 위해 앞서와 마찬가지로 의미적 연관도를 이용한다. 댓글 집합 C에 속하는 댓글 C_k와 의미적으로 가장 가까운 단어 $\hat{t}$는 식(3)과 같이 최대 의미적 연관도를 가지는 t_i이며, $w(\hat{t_i}, c_k)$는 식 (1)로 구할 수 있다.

$$w(\hat{t}, c_k) = \max_i w(t_i, c_k)$$
$$where\ t_i \in T, c_k \in C \qquad (3)$$

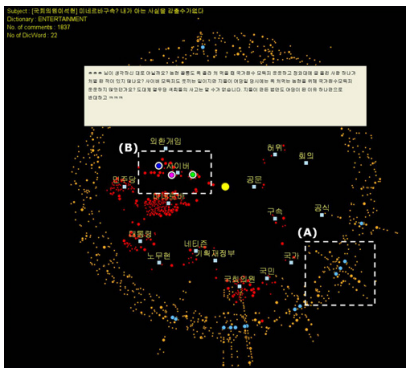
그림 3은 의미적 연관도에 따라 배치된 댓글들을 보여준다. 여기서, t_i는 사전에 정의된 i번째 단어이고, $c_{i,j}$는 단어 t_i에 속하는 j번째 댓글이다. 그림에서 t_i에 가까울수록 의미적 연관도가 높음을 의미하고, $c_{i,j}$의 크기는 댓글의 글자 수에 비례한다. 여기서, $d_{i,S}$는 $dist(t_i, S)$이고, $d_{i,j}$는 $dist(t_i, c_{i,j})$를 나타낸다.



(그림 3) 댓글의 배치

사전에 정의된 단어를 하나도 포함하지 않는 댓글은 단어로 분류되지 않고 댓글 작성자 ID별로 화면의 주변에 배치된다. 동일한 작성자가 올린 댓글의 경우 직선상에 배치하여 많은 댓글을 작성한 ID를 쉽게 식별할 수 있다.

그림 4는 TRIB으로 시각화 된 결과를 보여준다. 그림에서 붉은 원은 사전의 단어에 속한 댓글이며, 주변에 위치한 노란 색 원들은 단어에 속하지 않고 작성자 ID에 따라 배치된 댓글을 나타낸다. 특히 동일한 작성자가 많은 댓글을 올린 경우는 그림 4 (A)와 같이 표현되어 시각화 화면에서 쉽게 식별할 수 있다. 서로 논쟁을 하거나 앞의 댓글에 답하는 댓글을 작성할 경우 시간적으로 서로 인접한 경우가 많다. TRIB에서는 댓글의 순차적 검색을 돕기 위해 그림 4 (B)에서 보이는 것처럼 선택한 댓글은 보라색으로, 그 댓글의 앞과 뒤의 댓글은 녹색과 파란색으로 각각 표시해 사용자가 쉽게 찾을 수 있도록 해준다.



(그림 4) TRIB의 시각화

4. 실험 결과

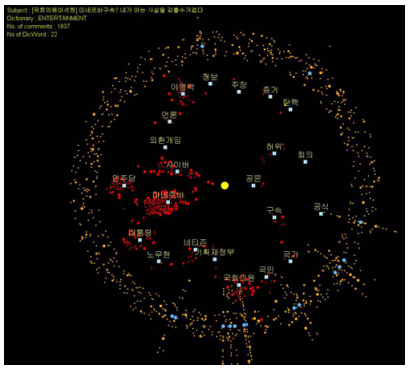
TRIB은 C#과 시각화 모듈을 위해 Processing으로 구현되었다. 본 논문에서의 실험은 많은 수의 댓글에 대한 분류 및 시각화 성능을 보이기 위해 댓글의 수가 1,000개 이상인 기사를 대상으로 하였으며 실험에 사용된 기사와 댓글은 인터넷 포털 사이트 'Daum'에서 운영하는 온라인 토론 게시판인 'AGORA'에서 수집되었다. 본 논문에서는 사용자 정의 사전을 정치와 연예에 관련된 두 가지로 정의하고 각각을 $PoliDic$과 $EnterDic$이라고 한다. 두 사전은 각각의 주제에 맞는 22개의 단어들로 구성되었다. 실험으로 정치관련 기사, 연예관련 기사, 그리고 일반 기사로 나누어 두 사전을 각각 적용하였다.

<표 2> 실험 기사 집합

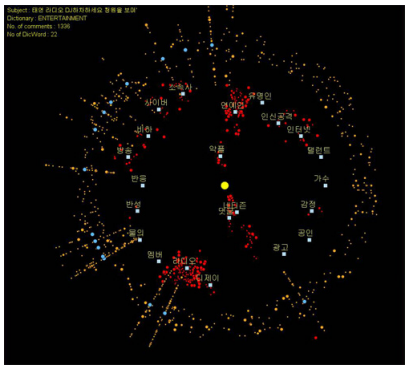
기사명	주제	댓글수	작성일
S_P	정치	1,837	09.01.10
S_E	연예	1,836	09.01.18
S_G	일반	1,838	09.01.15

표 2는 실험에 사용된 기사들의 주제 및 댓글 수를 정리한 것이다.

그림 5는 S_P와 S_E에 각각의 주제에 맞는 사전을 적용하여 시각화 한 결과를 보여준다. 두 기사 모두 단어의 주위로 많은 댓글들이 분류된 것을 볼 수 있다.



(a) S_P에 $PoliDic$을 적용한 결과

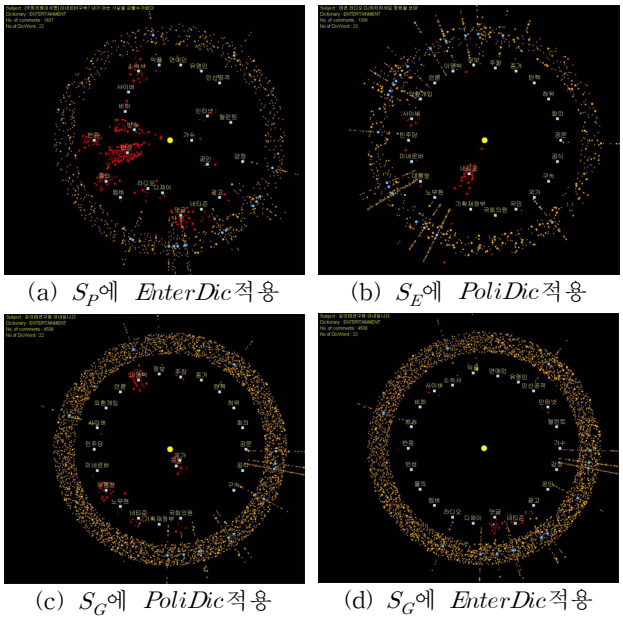


(b) S_E에 $EnterDic$을 적용한 결과

(그림 5) 기사의 주제에 맞는 사전을 사용한 시각화 결과

그림 6은 S_P와 S_E 그리고 S_G에 서로 주제와 관련 없는 사전을 적용하여 시각화 한 결과이다. S_G의 경우는 $PoliDic$과 $EnterDic$ 두 가지 사전 모두를 적용하였다. 그림 5에서와는 달리 단어로 분류되지 않고 ID 별로 분류

되어 주변에 나타나는 댓글들이 많음을 볼 수 있다.



(a) S_P에 $EnterDic$적용 (b) S_E에 $PoliDic$적용

(c) S_G에 $PoliDic$적용 (d) S_G에 $EnterDic$적용

(그림 6) 주제에 맞지 않는 사전을 적용하여 시각화 한 결과

<표 3> 사전별 댓글 분류 정도

기사명	댓글수	$PoliDic$		$EnterDic$	
		단어분류	비율	단어분류	비율
S_P	1,837	473	26%	134	7%
S_E	1,336	67	5%	335	25%
S_G	1,838	120	6.5%	26	1.4%

표 3은 주제별 기사와 사전의 적용에 따른 댓글 분류 정도를 정리한 것이다. 위의 실험을 통해 주제별 사용자 정의 사전의 사용으로 기사의 댓글들이 내용에 따라 효율적으로 분류되고 시각화됨을 알 수 있다.

TRIB을 통한 시각화 결과에서 우리는 흥미로운 사실을 발견하였다. 연예 관련 기사의 경우 정치나 일반적인 기사에 비해 ID로 분류된 댓글들이 직선으로 나타나는 경우가 많은 것을 볼 수 있다. 이것은 같은 기사에 반복적으로 댓글을 올리는 작성자가 많은 것을 의미한다. 비교적 연예 관련 기사의 검색과 댓글 작성의 연령대가 낮은 것을 고려한다면 젊은 인터넷 사용자들의 경우 인터넷 공간에서 더 적극적으로 의사 표현을 한다고 볼 수 있다.

<표 4> 댓글 작성이 많은 상위 10개 ID의 평균 댓글 수

기사명	평균 댓글 수	표준편차
S_P	28.9	7.63
S_E	40.9	18.68
S_G	30.0	9.75

표 4는 실험에 사용된 기사의 댓글 작성자 중 댓글을 많이 작성한 상위 10개의 ID가 작성한 댓글 수에 대한 통계를 보여준다.

5. 결론

본 논문에서는 인터넷 뉴스나 블로그 기사에 달린 많은 수의 댓글들을 내용에 따라 분류하고 시각화하는 시스템인 TRIB을 제안하였다. TRIB은 하나의 기사에 달린 댓글들에 대한 전체적인 개관을 제시한다. 사용자의 관심 주제에 맞는 단어 사전을 정의하고 이것을 댓글의 분류 기준으로 사용함으로써 댓글의 내용에 따른 검색이 가능하다. 또한 작성 시간에 따른 순차적 접근도 가능하므로 논쟁의 경우와 같이 서로 주고받는 형태의 댓글도 쉽게 찾아 볼 수 있다.

최근의 인터넷 사용자들은 인터넷 뉴스나 블로그 등을 통해 게시물들을 단순히 읽는 것에서 벗어나 댓글을 통해 적극적으로 자신의 의견을 피력하거나 다른 사람의 의견을 살피고 있다. 따라서 현재 이슈가 되는 기사들의 경우 수백에서 수천 개의 댓글을 가지기도 한다. 그중에서는 유용한 정보를 포함하는 댓글도 있는 반면 광고나 비속어등과 같이 불필요한 댓글들도 많이 포함되어 있다. TRIB은 내용에 따른 분류 및 시각화가 가능하므로 주제에 맞는 댓글뿐만 아니라 광고성 댓글이나 비속어 등이 많이 포함된 댓글도 차단할 수 있을 것으로 기대된다.

참고문헌

[1] 심재민, 조찬형, 양효진, 안인회, 나은아, “웹2.0 시대의 네티즌 인터넷 이용 현황”, *2006년 인터넷이슈심층조사 보고서*, 한국인터넷진흥원, 2006

[2] 김은미, 선유화, “댓글에 대한 노출이 뉴스 수용에 미치는 효과,” 한국언론학보, pp. 33-64, 2006.

[3] T. Nguyen and J.Zhang. "A novel visualization model for web search results." IEEE trans. Vis. Comput. Graph, 12(5):981-988, 2006.

[4] Harris J., and Kamvar S., We Feel Fine. http://www.wefeelfine.org, 2006.

[5] BBC, Spectrum, http://www.bbc.co.uk/white/spectrum.shtml.

[6] Indratmo, J., and Gutwin, C. “Exploring blog archives with interactive visualization”. In Proceedings of the Working Conference on Advanced Visual Interfaces, pages:39-46, 2008

[7] Y. Takama, A. Matsumura, and T. Kajinami. Visualization of News Distribution in Blog Space. In Proceedings of the 2006 IEEE/WIC/ACM international conference on Web Intelligence and Intelligent Agent Technology, pages 413-416, 2006.

[8] M. Gregory, D. Payne, D. McColgin, N. Cramer, and D. Love. Visual analysis of weblog content. In International Conference on Weblogs and Social Media, Boulder, 2007.