DOI: 10.3745/KIPSTD.2009.16D.5.817

TRIB: 블로그 댓글 분류 및 시각화 시스템

이 윤 정[†] · 지 정 훈[††] · 우 균[†††] · 조 환 규[††††]

요 약

최근 들어 블로그나 인터넷 게시판 등은 사람들의 정보 공유나 의견 교환의 중요한 매체가 되고 있으며, 많은 수의 블로그들이 사회적 문제들을 반영하고 있다. 온라인 커뮤니티에서 많은 사용자들은 댓글을 통해 인터넷 뉴스나 블로그 게시물에 대한 자신의 의견을 적극적으로 표현하고 있다. 블로그 사용이 활발해짐에 따라 수만개 이상의 댓글들이 등록되는 블로그들도 쉽게 찾을 수 있다. 대부분의 블로그나 인터넷 포털 사이트의 경우 게시물이나 댓글들을 순차적인 목록 형태로 제공하므로 자신이 원하는 내용의 댓글을 검색하거나 전체 댓글에 대한 전반적인 파악이 힘들다. 본 논문에서는 게시물에 달린 많은 수의 댓글들을 분류하고, 이를 시각화 하는 시스템인 TRIB (Telescope for Responding comments for Internet Blog)를 제안한다. TRIB는 미리 정의된 사용자 정의 사전을 이용하여 댓글을 내용에 따라 분류하여 시각화한다. 또한, 사용자들의 관심과 흥미를 고려한 개인화 된 뷰를 제공한다. TRIB의 유용성을 보이기 위해서 1,000개 이상의 댓글을 가진 인터넷 게시물들을 대상으로 한 실험을 통해 TRIB 시스템의 댓글 분류와 시각화 성능을 보인다.

키워드 : 웹 블로그, 댓글 시각화, 댓글 분류 시스템

TRIB : A Clustering and Visualization System for Responding Comments on Blogs

Yun Jung Lee[†] · Jung Hoon Ji[††] · Gyun Woo[†††] · Hwan Gue Cho[††††]

ABSTRACT

In recent years, Weblog has become the most typical social media for citizens to share their opinions. And, many Weblogs reflect several social issues. There are many internet users who actively express their opinions for internet news or Weblog articles through the replying comments on online community. Hence, we can easily find internet blogs including more than 10 thousand replying comments. It is hard to search and explore useful messages on weblogs since most of weblog systems show articles and their comments to the form of sequential list. In this paper, we propose a visualizing and clustering system called TRIB (Telescope for Responding comments for Internet Blog) for a large set of responding comments for a Weblog article. TRIB clusters and visualizes the replying comments considering their contents using pre-defined user dictionary. Also, TRIB provides various personalized views considering the interests of users. To show the usefulness of TRIB, we conducted some experiments, concerning the clustering and visualizing capabilities of TRIB, with articles that have more than 1,000 comments.

Keywords : Weblog, Comment Visualization, Comment Clustering System

1. 서 론

최근 들어 인터넷 게시판이나 개인 블로그 등은 온라인 상에서 사람들의 정보 공유나 의견 교환의 중요한 매체가 되고 있으며 온라인 커뮤니티를 형성하며, 현재 사회적으로 이슈가 되는 여러 문제들을 반영하고 있다. 온라인 커뮤니티에서 사용자들은 게시물을 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 답글이나 덧글과 같은 용어로도 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 있

※ 이 논문은 부산대학교 자유과제 학술연구비(2년)에 의하여 연구되었음
† 정 회 원 : 부산대학교 U-Port 정보기술사업단 박사후연구원
†† 준 회 원 : 부산대학교 컴퓨터공학과 박사과정
††† 종신회원 : 부산대학교 컴퓨터공학과 부교수(교신저자)
†††† 정 회 원 : 부산대학교 컴퓨터공학과 교수
논문접수 : 2009년 4월 20일
수 정 일 : 1차 2009년 7월 24일
심사완료 : 2009년 7월 24일

다[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 달린 댓글의 수는 기사의 중요도나 관심 정도에 따라 다르긴 하지만 적게는 수백 개에서 많게는 몇 만개 이상이 되기도 한다. 국내의 대표적인 온라인 커뮤니티로 인터넷 포털 사이트인 다음(Daum)에서 운영하는 아고라(AGORA)를 들 수 있다. 다음의 아고라에서는 수많은 네티즌들이 모여 정보를 공유하고 토론을 벌이고 있으며 하루에도 수천 건 이상의 게시물들이 올라오고, 각 게시물에는 관심 정도에 따라 수백에서 수만 개 이상의 댓글이 달리기도 한다.

최근 댓글 이용자가 늘어남에 따라 익명성을 이용하여 게시물의 내용과 관련 없는 광고성 글이나 비속어 등이 사용된 악성 댓글들도 다수 포함되어 있어 사회 문제가 되기도 한다. 현재 대부분의 온라인 게시판이나 블로그에서는 목록 형태로 데이터를 제공하고 있어 많은 수의 게시물과 댓글들을 효율적으로 검색하기가 어렵다. 최근 게시물이나 댓글의 태그(tag) 정보를 이용하여 분류나 검색 서비스를 제공하고 있으나 태그 정보를 신뢰하기 어렵고 태그가 없는 게시물도 많아 효율적인 검색이 어렵다. 게시물의 경우는 태그 외에도 제목, 내용, 작성자 등과 같은 정보를 통해 검색이나 정렬이 가능하나 댓글의 경우는 내용에 따른 검색이나 댓글간의 연관성을 파악하는 것과 같은 2차적 데이터 처리는 제공되고 있지 않다. 앞서 언급된 것과 같이 온라인 커뮤니티에서 의견 수렴과 정보 공유의 도구로 댓글이 유용하게 활용되는 만큼 효율적인 댓글 검색과 댓글간의 연관관계를 파악할 수 있는 방법이 필요할 것이다.

따라서 본 논문에서는 인터넷 뉴스나 블로그 게시물에 달린 많은 수의 댓글들을 사용자 정의 사전을 통해 내용에 따라 분류하고 이를 시각화하는 시스템인 TRIB를 제안한다. TRIB에서는 화면 중심에 게시물을 배치하고 게시물의 내용과 연관 정도에 따라 사용자 정의 사전의 단어들을 그 주변에 배치한다. 그 게시물에 달린 댓글들은 자신과 가장 의미적으로 연관도가 높은 단어에 속하게 되어 그 단어의 주변에 배치된다. TRIB의 화면 구성은 Nguyen[3]의 연구에서와 마찬가지로 태양계와 유사하다. TRIB는 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

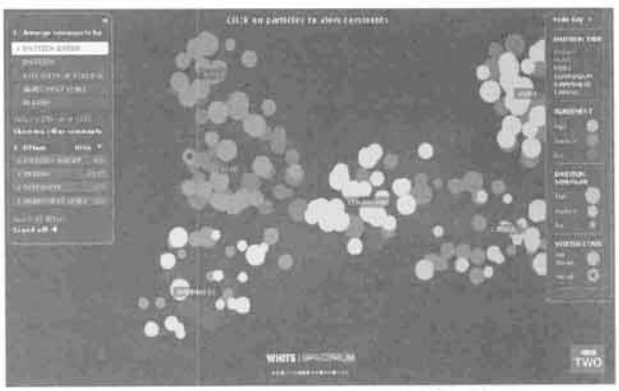
Harris[4] 등은 블로그 시각화 방법인 "We feel fine"이라는 시스템을 개발하였다. 일정 시간마다 전 세계에서 게시되는 블로그 게시물들을 수집하고 게시물에 포함된 감정 표현 문장들을 분석하여 행복(happy), 슬픔(sad), 우울(depressed)과 같은 감정 상태로 분류한다. (그림 1)은 "We feel fine" 시스템의 시각화 결과를 보여준다.

이 시스템에서 각각의 감정 상태들은 색상이 다른 도형으로 표현되어 시각화된다. "We feel fine" 시스템은 많은 블로그 게시물들을 표현하고 있으나 어떤 블로그에서 게시되었는지 혹은 게시물들의 앞, 뒤 연결을 알 수 없다는 단점이 있다.

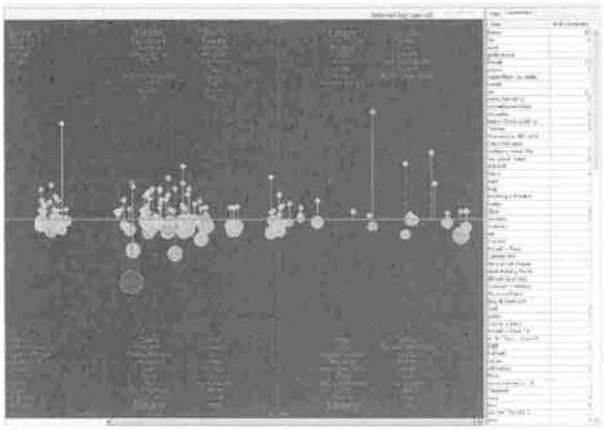
이것과 유사하게 BBC에서는 뉴스에 대한 댓글을 시각화하는 시스템인 Spectrum을 개발하였다[5]. BBC 2's White 시즌 중 토론을 조사하여 감정, 지역, 성별 등에 따라 댓글을 클러스터링하고 이를 시각화한다. (그림 2)는 Spectrum의 시각화 결과를 보여준다.

Spectrum에서 각 댓글들은 감정 별로 분류되어 서로 다른 색의 원으로 시각화되며, 감정과 지역, 성별 등과 같이 그룹화 할 기준을 선택할 수 있는 사용자 인터페이스를 제공하고 있어 원하는 기준으로 댓글들을 필터링할 수 있고

(그림 1) "We feel fine"의 시각화 결과[4]



(그림 2) Spectrum의 시각화 화면[5]

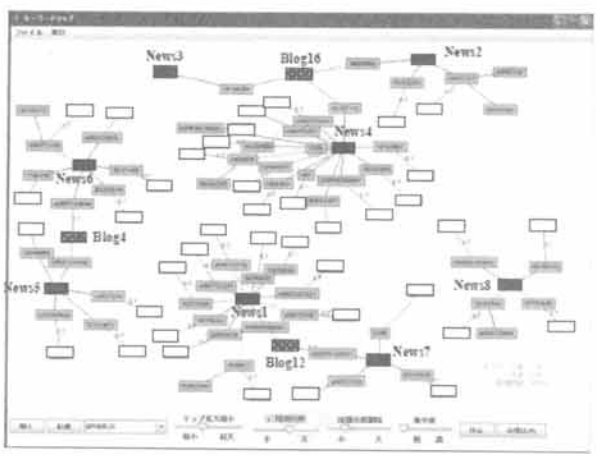


(그림 3) iBlogVis의 블로그 시각화 화면[6]

움직이는 입자를 클릭하면 토론에서 사용된 댓글을 볼 수 있다. 그러나 "We feel fine" 시스템과 마찬가지로 Spectrum도 블로그나 뉴스 카테고리 내에서 한 항목에 대한 앞, 뒤 순서나 연결 상태를 알 수 없으며, 감정 표현 상태를 제외하고는 댓글의 내용과는 관계없는 성별, 나이, 지역과 같은 작성자의 환경에 따른 분류만 가능하므로 실제 관심 이슈에 대한 내용을 담고 있는 글을 검색하는 데에는 어려움이 있다.

Indratmo[6] 등은 블로그 시각화 도구인 iBlogVis 시스템을 제안하였다. iBlogVis에서는 블로그 내의 게시물들을 게시된 시간에 따라 수평인 시간 축의 위쪽에 배치하고 아래쪽에는 해당 게시물에 달린 댓글들의 개수나 글자 수를 고려하여 시간축의 아래쪽에 배치한다. 그림 3은 iBlogVis의 시각화 결과를 보여준다.

(그림 3)에서 마름모 형태는 게시물을 나타내고 시간 축과 연결된 선의 길이는 게시물의 글자 수에 비례한다. 마찬가지로 시간 축 아래쪽의 원들은 같은 수직선상에 놓인 게시물에 달린 댓글을 나타내며, 원의 크기는 댓글의 개수에 비례하며 시간 축과 연결된 선의 길이는 댓글의 총 글자 수에 비례한다. 시각화된 화면에서 실제 게시물의 내용을 알 수 없으며 단지 블로그에 대한 개요만 제공한다. iBlogVis에



(그림 4) 블로그 공간의 시각화[7]

〈표 1〉 기존 시스템 비교

시스템명	시각화 대상	분류기능	내용 기반 분류
We Feel Fine[4]	블로그 게시물	Y	N
Spectrum[5]	댓글	Y	N
iBlogVis[6]	블로그	Y	N
Takama's[7]	블로그 공간	N	N
TRIB	댓글	Y	Y

서는 블로그 내의 게시물과 댓글의 전체적인 현황은 파악할 수 있으나 리스트로 제공되는 것과 마찬가지로 댓글의 검색이나 게시물과의 의미적 관계 등은 파악할 수 없다.

Takama[7] 등은 블로그 공간(blog space)에서 뉴스 기사와 블로그 게시물 그리고 블로그 사이트의 분포를 시각화하는 방법을 제안하였다. 이 방법에서는 뉴스 기사의 중요도와 블로그 링크를 사용하여 블로그 사이트와 게시물들을 시각화한다. 게시물의 중요도는 조회 수로 계산된다. 그림 4는 시각화된 블로그 공간을 보여준다.

이외에도 블로그 게시물의 태그 정보나 트랙백과 같은 다양한 정보를 활용한 게시물 분류와 시각화에 대한 연구들이 현재 활발히 진행되고 있다[8-11]. <표 1>은 대표적인 온라인 커뮤니티인 블로그와 관련된 기존의 시각화 시스템들의 기능을 비교한 것이다.

<표 1>에서 볼 수 있는 바와 같이 Spectrum을 제외한 모든 시각화 시스템이 블로그 게시물이나 블로그 공간만을 시각화 대상으로 하고 있으며, 댓글에 대한 검색이나 시각화에 대한 연구는 찾아보기 어려웠다.

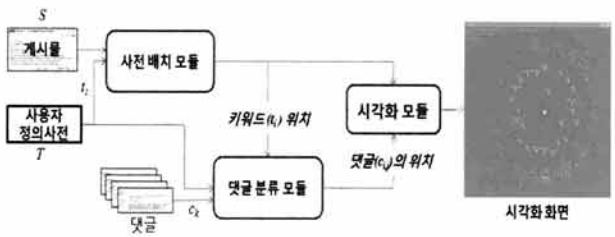
본 논문에서 제안하는 TRIB는 블로그 게시물에 추가된 댓글을 시각화 대상으로 삼는다. Spectrum과 같은 다른 시각화 시스템들은 댓글을 대상으로 하고 있긴 하지만 내용을 기반으로 한 분류를 수행하고 있지 않다는 점에서 한계가 있다[12-14]. 본 논문에서 제시하는 TRIB는 댓글을 대상으로 시각화를 수행하면서도 댓글 내용을 기반으로 하여 댓글들을 분류하고 있다는 점에서 기존 시스템과 다른 새로운 방향으로 시각화를 수행한다고 할 수 있다.

3. 제안 시스템

3.1 TRIB의 개요

TRIB는 인터넷 게시물에 달린 많은 양의 댓글을 사용자 정의 사전을 이용하여 내용에 따라 분류하고 이를 시각화하는 시스템이다. TRIB의 시스템 구성은 (그림 5)와 같다.

TRIB는 사전 배치 모듈, 댓글 분류 모듈 그리고 시각화 모듈과 같이 세 부분으로 구성된다. (그림 5)에서 S는 게시물을 나타내고, c_k는 그 게시물에 달린 댓글 집합 C에 속한 k번째 댓글을 의미한다. 사용자 정의 사전 T는 키워드 집합으로, T에 속한 각각의 t_i 는 댓글 분류 시 키 속성으



(그림 5) TRIB의 시스템 구성도

로 사용된다.

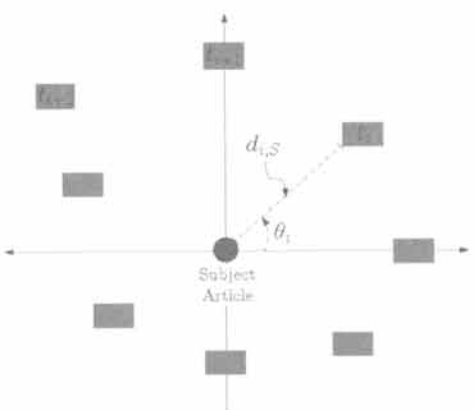
TRIB는 Windows 운영체 기반의 PC에서 C#과 Processing으로 구현되었다. 시각화 모듈을 제외한 모든 부분(사전배치 모듈과 댓글분류 모듈)은 C#으로 구현되었으며, 시각화 모듈은 Processing으로 구현되었다. Processing은 비주얼 컨텍스트 생성을 위한 오픈 소스 프로그래밍 언어이다[15].

3.2 사용자 정의 사전의 키워드 배치

많은 수의 댓글에 대해 의미 있는 뷰를 생성하기 위해서 사용자 사전의 정의가 중요하다. 예를 들어, 정치적인 이슈에 관련된 댓글을 분류하기 위해서 사전은 국회, 대통령, 한나라당, 민주당 등과 같은 단어를 포함할 수 있다. 이 밖에도 사용자의 관심 주제별로 정치, 경제, 연예 등 여러 개의 사전을 생성할 수 있다. 사용자 정의 사전 T에 속한 키워드 t_i는 태양계와 비슷하게 중심에 위치한 게시물 S의 주변으로 배치된다. 이때, 각각의 키워드는 게시물과의 의미적 연관 정도에 따라 연관도가 높을수록 S와 가까운 곳에 배치된다. 사용자 정의 사전 T에 속하는 키워드 t_i와 게시물 S와의 의미적 연관도를 계산하기 위하여 본 논문에서는 식 (1)과 같이 정의하였다.

$$w(t_i,S) = f_i / \sum_{k=1}^{|T|} f_k, \quad \forall_{t_i} \in T \tag{1}$$

여기서, f_i는 t_i가 S에서 나타난 횟수이며, $|T|$는 T에 속하는 키워드의 개수이다. 사전에 정의된 키워드들과 게시물 내용과의 의미적 연관도가 구해지면 키워드 t_i는 게시물을 중심으로 방사형으로 배치된다. (그림 6)은 사전에 정의된 키워드들의 배치를 보여준다.



(그림 6) 사전에 정의된 키워드들의 배치

(그림 6)에서 키워드 t_i의 위치는 회전각 θ_i와 중심과의 거리 $d_{i,S}$로 정해지며, 회전각과 거리는 식 (2)와 같이 계산된다.

$$\begin{aligned} d_{i,S} &= R \cdot \exp(-c \cdot w_i) \\ \theta_i &= 2\pi \cdot i/|T| \end{aligned} \tag{2}$$

회전각 θ_i는 사전에서 키워드 t_i의 순서에 따라 수평에서 반시계 방향으로 등간격으로 구해지고, 중심과의 거리 $d_{i,S}$는 게시물과 키워드 t_i 간의 의미적 연관도인 w_i를 이용해 계산된다(식 (1)참고). R은 t_i가 위치될 수 있는 최대 거리이며, 상수 c는 중심으로의 밀집도를 조절하는 인수이다. 간단히 말해서 키워드가 게시물에 가까울수록 게시물에 자주 나타나는 키워드임을 의미한다.

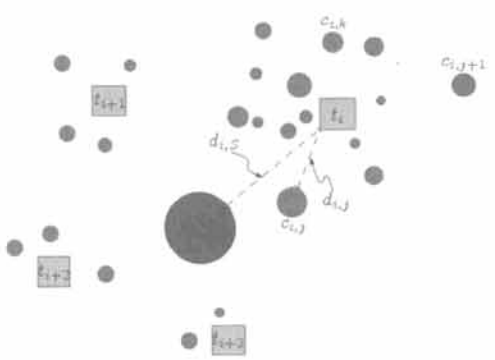
3.3 댓글의 배치

댓글들은 사전에 정의된 키워드 중 자신과 의미적으로 가장 가까운 키워드의 주변에 배치된다. 댓글과 가장 의미적으로 가까운 키워드를 찾기 위해 앞서와 마찬가지로 의미적 연관도를 이용한다. 댓글 집합 C에 속하는 댓글 c_k와 의미적으로 가장 가까운 키워드 $\hat{t}$는 식 (3)과 같이 최대 의미적 연관도를 가지는 t_i이며, $w(\hat{t_i}, c_k)$는 식 (1)로 구할 수 있다.

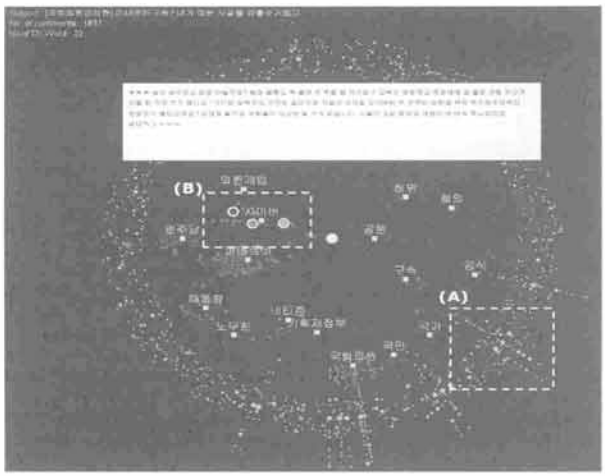
$$w(\hat{t_k}, c_k) = \max(w(t_i, c_k)), \quad where\ t_i \in T, c_k \in C \tag{3}$$

(그림 7)은 의미적 연관도에 따라 키워드 주변으로 배치된 댓글들을 보여준다. (그림 7)에서 $\{c_{i,j}\}$는 키워드 t_i에 속하는 댓글 집합을 의미한다. 그림에서 $c_{i,j}$는 $c_{i,j+1}$보다 t_i에 더 가깝다. 이것은 $c_{i,j}$가 $c_{i,j+1}$보다 t_i와 의미적 연관도가 더 높음을 의미한다. 또한 댓글을 표현하는 원의 크기는 해당 댓글의 글자 수에 비례한다.

사전에 정의된 키워드를 하나도 포함하지 않는 댓글은 키워드로 분류되지 않고 댓글 작성자 id별로 화면의 주변에 배치된다. 동일한 작성자가 올린 댓글의 경우 직선상에 배



(그림 7) 댓글의 배치



(그림 8) TRIB의 시각화

치하여 많은 댓글을 작성한 id를 쉽게 식별할 수 있다.

(그림 8)은 아고라에서 수집한 게시물 중 1,837 개의 댓글을 가지는 게시물을 TRIB를 통해 시각화한 결과 화면이다. 그림에서 크기가 다른 붉은 색과 주황색의 원들은 게시물에 달린 댓글을 나타낸다.

사용자가 원을 클릭하면 그림에서와 같이 팝업창을 통해 해당 댓글의 내용을 확인할 수 있다. (그림 8)에서 붉은 원은 사전의 키워드에 속한 댓글이며, 주변에 위치한 주황색 원들은 어느 키워드와도 의미적 연관이 없어 작성자 ID에 따라 배치된 댓글을 나타낸다. 특히 동일한 작성자가 많은 댓글을 올린 경우는 (그림 8)의 (A) 부분과 같이 일직선상에 많은 댓글들이 표현되어 시각화 화면에서 쉽게 식별할 수 있다.

(그림 9)는 (그림 8)(A) 부분을 확대한 것으로 파란 원은 많은 수의 댓글을 작성자 id를 나타내고 이것을 클릭하면 해당 id를 확인할 수 있다.

사용자가 서로 논쟁을 하거나 앞의 댓글에 답하는 댓글을 작성할 경우 시간적으로 서로 인접한 경우가 많다. TRIB에서는 (그림 8)의 (A) 부분과 같이 선택된 댓글의 앞과 뒤의 댓글을 다른 색상의 이중 원으로 표시함으로써 댓글의 순차적 검색을 돕는다. (그림 10)은 (그림 8)의 (B) 부분을 확대한 것으로 선택한 댓글은 보라색으로, 그 댓글의 앞과 뒤의 댓글은 녹색과 파란색으로 각각 표시해 사용자가 쉽게 찾을 수 있도록 해준다.

(그림 9) 댓글 작성자 ID 별 분류 결과

(그림 10) 댓글의 순차 검색

4. 실험 결과

본 논문에서의 실험은 TRIB의 사전에 따른 댓글 분류와 시각화 성능을 보이기 위해 실험을 수행하였다. 실험에 사용된 게시물과 댓글은 인터넷 포털 사이트 '다음'에서 운영하는 온라인 토론 게시판인 '아고라'에서 수집되었다. 수집된 게시물 중에서 댓글 수가 1,000개 미만인 것은 제외하였다.

4.1 실험 데이터

본 논문에서는 실험을 위해 2009년 1월 23일부터 2009년 1월 29일까지 일주일 동안 '아고라'에 게시되는 게시물을 수집하였다. 대량의 댓글에 대한 시각화 성능을 보이기 위해 주제에 관련 없이 조회 수가 많은 '토론 베스트'에 올라온 게시물들로 제한하였다. 수집된 게시물의 수는 <표 2>와 같다. 표 2에서 보듯이 날짜별로 차이를 보이지만 하루 평균 1,500개 이상의 게시물이 올라오는 것을 알 수 있다.

본 실험에서는 수집된 게시물 중에 정치, 연예, 일반 범주에 속한 게시물 중 댓글 수가 비슷한 게시물을 선택하였다. 실험 데이터는 <표 3>과 같다. <표 3>에서 S_P와 S_E는 각각 정치와 연예에 관련된 게시물이며, S_G는 특정 주제가 없는 일반적인 게시물을 나타낸다. 세 가지 게시물 모두 1,000개 이상의 댓글을 가진다.

댓글 분류를 위해 사용되는 사용자 정의 사전은 정치와

〈표 2〉 '아고라' 게시판의 '토론베스트'에 등록된 날짜별 게시물 개수 및 평균 게시물 수

날짜	게시물 수
2009년 1월 23일	2,678
2009년 1월 24일	1,929
2009년 1월 25일	672
2009년 1월 26일	957
2009년 1월 27일	1,010
2009년 1월 28일	1,870
2009년 1월 29일	1,887
평균 게시물 수	1,572

〈표 3〉 실험 게시물

게시물	주제	댓글 수
S_P	정치	1,837
S_E	연예	1,336
S_G	일반	1,838

〈표 4〉 *PoliDic*에 정의된 키워드

공문	공식	구속	국가
국민	국회의원	기획재정부	네티즌
노무현	대통령	미네르바	민주당
사이버	외환개입	언론	이명박
정보	주장	증거	탄핵
허위	회의		

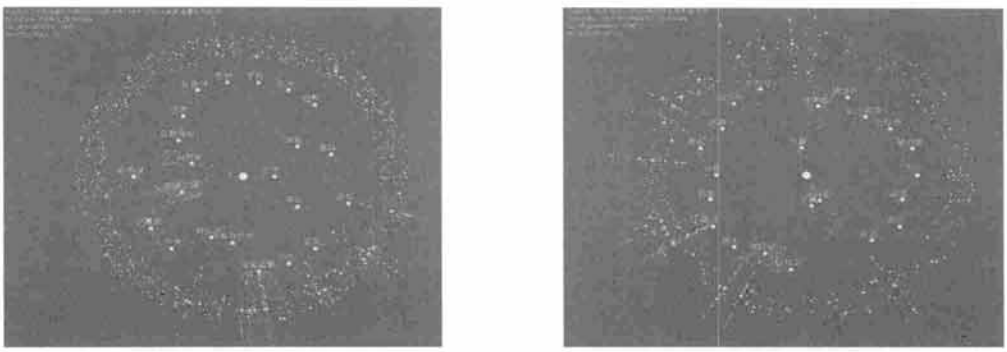
〈표 5〉 *EnterDic*에 정의된 키워드

가수	공인	감정	광고
네티즌	댓글	디제이	라디오
멤버	물의	반성	반응
방송	비하	사이버	소속사
악플	연예인	유명인	인신공격
인터넷	탤런트		

연예에 관련된 두 가지로 정의하고 각각을 *PoliDic*과 *EnterDic*이라 하고, 두 사전은 각각의 주제에 맞는 단어와 일반적으로 게시물에 많이 나타나는 단어 등을 포함해 22개의 키워드로 구성되었다. 두 사전에 정의된 키워드들은 <표 4>와 <표 5>에 나타나 있다.

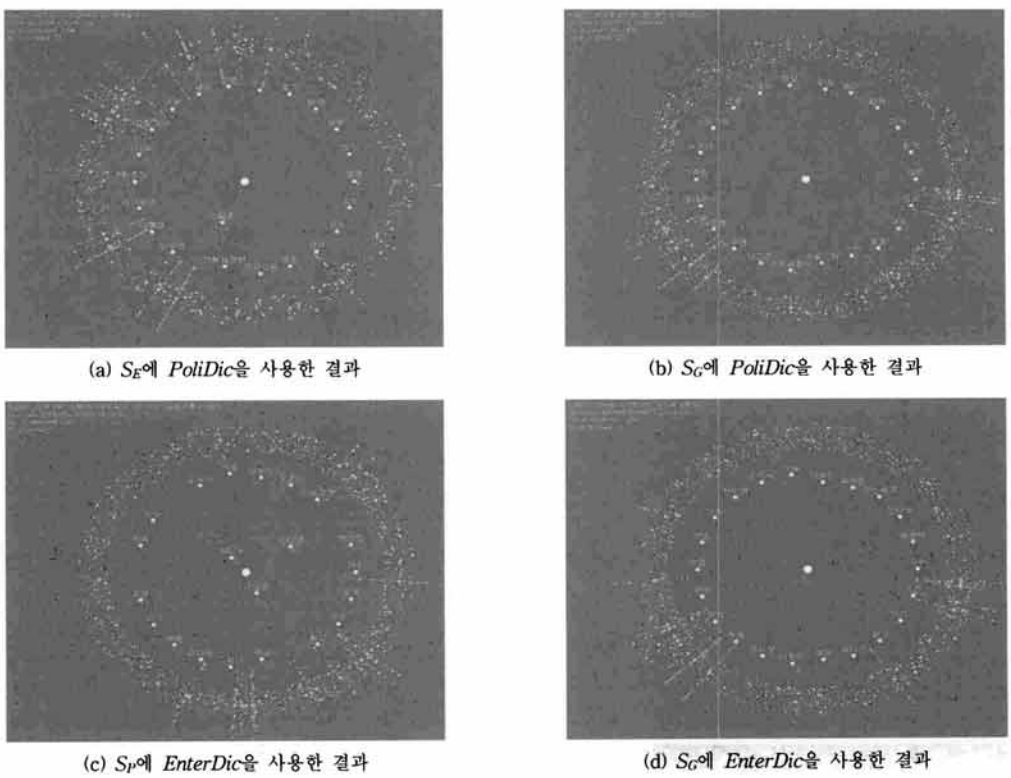
(그림 11)은 S_P와 S_E에 각각의 주제에 맞는 사전을 적용하여 시각화한 결과를 보여준다. (그림 11)(a)는 정치 관련 게시물인 S_P에 *PoliDic*을 이용하여 시각화한 결과이며, (b)는 연예 관련 게시물인 S_E에 *EnterDic*을 이용하여 시각화한 결과이다. 두 경우 모두 키워드의 주위로 많은 댓글들이 분류된 것을 볼 수 있다.

(그림 12)는 S_P, S_E와 S_G에 게시물의 주제와 관련 없는 사



(a) S_P에 *PoliDic*을 적용한 결과 (b) S_E에 *EnterDic*을 적용한 결과

(그림 11) 게시물의 주제에 맞는 사전을 사용한 시각화 결과



(a) S_E에 *PoliDic*을 사용한 결과 (b) S_G에 *PoliDic*을 사용한 결과

(c) S_P에 *EnterDic*을 사용한 결과 (d) S_G에 *EnterDic*을 사용한 결과

(그림 12) 게시물의 주제에 맞지 않는 사전을 사용한 시각화 결과

전을 적용하여 시각화한 결과를 보여준다. (그림 12)의 (a)와 (b)는 S_P와 S_G에 *PoliDic*을 사용하여 시각화한 결과이다. 그리고 (c)와 (d)는 S_P와 S_G에 *EnterDic*을 사용하여 시각화한 결과이다.

4가지 결과 모두 (그림 11)에서와는 대조적으로 키워드 주변으로 배치되지 않고 댓글 작성자 id 별로 분류된 댓글이 많음을 볼 수 있다.

<표 6>은 주제별 게시물과 사전의 적용에 따른 댓글 분류 정도를 정리한 것이다. 위의 실험을 통해 주제별 사용자 정의 사전의 사용으로 게시물의 댓글들이 내용에 따라 효율적으로 분류되고 시각화됨을 알 수 있다.

TRIB를 통한 시각화 결과에서 우리는 흥미로운 사실을 발견하였다. 연예 관련 게시물의 경우 정치나 일반적인 게시물에 비해 id로 분류된 댓글들이 직선으로 나타나는 경우가 많은 것을 볼 수 있다. 이것은 같은 게시물에 반복적으로 댓글을 올리는 작성자가 많은 것을 의미한다. <표 7>은 실험에 사용된 게시물의 댓글 작성자 중 댓글을 많이 작성한 상위 10개의 id가 작성한 댓글 수에 대한 통계를 보여준다.

비교적 연예 관련 게시물의 검색과 댓글 작성의 연령대가 낮은 것을 고려한다면 젊은 인터넷 사용자들의 경우 인터넷 공간에서 더 적극적으로 의사 표현을 한다고 볼 수 있다.

〈표 6〉 사전별 댓글 분류 정도

게시물	댓글 수	*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 7〉 댓글 작성이 많은 상위 10개 ID의 평균 댓글 수

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

5. 결 론

인터넷 게시판이나 블로그 등은 온라인상에서 사람들의 정보 공유나 의견 교환의 중요한 매체가 되고 있으며, 사용자들은 게시물을 읽고 정보를 얻는 것뿐만 아니라 댓글을 통해 타인의 의견을 살피거나 자신의 생각을 좀 더 적극적으로 나타내고 있다. 이처럼 참여형 인터넷 사용자가 증가함에 따라 각종 블로그나 인터넷 게시판 등에는 엄청난 수의 게시물들과 댓글들이 게시되고 있다. 그러나 대부분의 블로그나 인터넷 게시판의 경우 게시물이나 댓글들을 목록으로 제공하고 있어 게시물에 대한 전체적인 개관을 파악하기 어렵고, 작성자, 제목 등 몇 가지 기준으로의 검색이나 정렬을 제외하고는 원하는 정보를 검색하기 어려운 실정이다. 특히 게시물에 달린 댓글의 경우는 이러한 정렬이나 검색조차도 지원되지 않고 있다.

기존의 블로그 관련 시각화 연구들은 주로 블로그 공간이나 많은 양의 게시물을 하나의 화면에 보여주기 위한 연구에 치중에 있는 반면에 본 논문에서는 게시물에 달린 많은 수의 댓글들을 분류하고 이를 시각화하는 시스템인 TRIB를 제안하였다. TRIB는 내용을 기반으로 하여 댓글을 분류할 수 있다. 사용자는 자신이 관심 있는 단어들을 이용하여 사전을 정의할 수 있으며, 사전의 단어들은 댓글 분류의 키워드로 사용된다. 또한 정의된 단어와 관계없는 댓글의 경우 작성자 별로 따로 분류되어 반복적으로 많은 댓글을 작성하는 작성자도 쉽게 구분할 수 있다. 또한 TRIB는 분류된 댓글들을 하나의 뷰로 시각화하고, 댓글 접근을 위한 사용자 인터페이스를 제공한다. 시각화 화면을 통해 사용자들은 게시물에 달린 많은 수의 댓글들에 대한 전체적인 개관을 파악할 수 있으며, 자신이 원하는 댓글을 손쉽게 읽을 수 있다. 시각화된 화면을 통한 댓글의 임의 접근 뿐만 아니라 댓글이 달린 시간 순서에 따른 순차적 접근도 가능하므로 논쟁의 경우와 같이 서로 주고받는 형태의 댓글도 쉽게 찾아 볼 수 있다.

향후 연구로서 악성 댓글을 차단하도록 TRIB를 확장하는 것을 생각해 볼 수 있다. 본 논문에서 제안한 댓글 시각화 시스템 TRIB는 내용에 따라 분류 및 시각화를 수행하고 있으므로 TRIB를 확장하면 악성 댓글을 차단할 수 있을 것으로 기대된다. 최근 광고성 댓글이나 비속어 등이 많이 포함된 댓글이 사회적으로 문제가 되고 있는데, 댓글 내용을 검색하여 비속어 등을 검출해 내고 이를 차단하도록 TRIB를 확장할 수 있을 것이다.

참 고 문 헌

[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 transaction on Visualization and Computer Graphics, Vol.12, No.5, pp.981-988, 2006.

[4] "We Feel Fine", http://www.wefeelfine.org

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

[6] J. Indratmo and C. Gutwin, "Exploring blog archives with interactive visualization," In Proceedings of the Working Conference on Advanced Visual Interfaces, pp.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, pp.413-416, 2006.
[8] S. Fujimura, K. Fujimura, and H. Okuda, "Blogosonomy: Autotagging any text using bloggers' knowledge," Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp.205-212, 2007.
[9] K. Fujumura, S. Fujimura, T. Matsubayash, T. Yamada, and H. Okuda, "Topigraphy: visualization for large-scale tag clouds," In WWW'08: Proceeding of the 17th international conference on World Wide Web, pp.1087-1088, 2008.
[10] J. Kim, K. Candan, and J. Tatemura, "CDIP: Collection-Driven, yet Individuality-Preserving Automated Blog Tagging," ICSC2007, pp.87-94, 2007.
[11] O. Kaser and D. Lemire, "Tag-Cloud Drawing: Algorithms for Cloud Visualization," WWW'07: 16th International World Wide Web Conference, 2007.
[12] G. Mishne and M. de Rijke, "MoodViews: Tools for blog mood analysis," In AAAI2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW2006), pp.153-154, 2006.
[13] C. Yang, K. Lin, and H. Chen, "Emotion Classification Using Web Blog Corpora," In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp.275-278, 2007.
[14] Y. Jung, Y. Choi, and S. Myaeng. "Determining Mood for a Blog by Combining Multiple Sources of Evidence," In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp.271-274, 2007.
[15] "Processing 1.0", http://www.processing.org

이 윤 정

e-mail : leeyj01@pusan.ac.kr
1995년 부경대학교 전자계산학과(학사)
1999년 부경대학교 전산정보학과 (이학석사)
2008년 부경대학교 전자계산학과 (공학박사)
2008년~현 재 부산대학교 U-Port 정보기술사업단 박사후연구원
관심분야 : 얼굴 애니메이션, 웹 컨텐츠 시각화

지 정 훈

e-mail : jhji@pusan.ac.kr
2003년 경성대학교 컴퓨터공학과(학사)
2005년 경성대학교 컴퓨터공학과 (공학석사)
2005년~현 재 부산대학교 컴퓨터공학과 박사과정
관심분야 : 프로그래밍 언어 및 컴파일러, 프로그램 표절검사, 자바가상기계, 프로그램 시각화

우 균

e-mail : woogyun@pusan.ac.kr
1991년 한국과학기술원 전산학(학사)
1993년 한국과학기술원 전산학(공학석사)
2000년 한국과학기술원 전산학(공학박사)
2000년~2002년 동아대학교 컴퓨터공학과 전임강사
2002년~2004년 동아대학교 컴퓨터공학과 조교수
2004년~현 재 부산대학교 컴퓨터공학과 부교수
관심분야 : 프로그래밍언어 및 컴파일러, 함수형 언어, 그리드컴퓨팅, 소프트웨어 메트릭, 프로그램 시각화

조 환 규

e-mail : hgcho@pusan.ac.kr
1984년 서울대학교 계산통계학과(학사)
1990년 한국과학기술원 전산학(공학박사)
1991년~현 재 부산대학교 컴퓨터공학과 · 교수
관심분야 : 알고리즘 이론, 생물정보학