인터넷 게시물의 댓글 분석

탁해성*, 조환규*
*부산대학교 정보컴퓨터공학부
e-mail : tok33@pusan.ac.kr, hgcho@pusan.ac.kr

Analysis for Comment of Internet Posts

Haesung Tak*, Hwan-Gue Cho*
*Dept. of Computer Science & Engineering, Pusan National University

요 약

최근 블로그나 인터넷 게시판과 같은 온라인 커뮤니티가 활발히 사용됨에 따라 댓글을 통해 자신의 의견을 적극적으로 나타내고자하는 이용자들이 계속해서 증가하고 있다. 실제 댓글 활동이 활발한 인터넷 게시판에서는 수천 개의 댓글이 달린 게시물도 심심치 않게 찾아볼 수 있다. 본 논문에서는 인터넷 게시판의 글의 정보와 댓글을 이용하여, 댓글의 확장과 조회 수와의 상관관계에 대해 알아보았다.

1. 서론

인터넷 게시판과, 온라인 카페, 커뮤니티, 포럼, 블로그, 그리고 최근에 각광 받고 있는 SNS에 이르기까지 다양한 인터넷 미디어들은 상호간의 의사소통이 가능한 매체들로 이루어져 있으며, 게시물 작성 및 열람, 댓글 달기, 채팅, 여론조사 참여, 본문 스크랩, 사진 올리기 등과 같은 행위를 포괄적으로 담고 있다.

댓글의 경우 해당 게시물에 대한 가장 적극적인 의사표현 형태로 볼 수 있으며, 게시물과 게시물 독자들과 상호작용할 수 있는 쉽고 효율적인 방법으로 사용되고 있다. 한 게시물에 추가되는 댓글의 수는 게시물의 내용이나 사용자들의 관심정도에 따라 수개에서 많게는 수천 개 이상으로 다양하게 나타난다. 흔히 많은 댓글이 달린 게시물은 사용자의 관심을 많이 받은 게시물로 간주되어 사용자들로부터 지속적인 관심을 받게 된다. 이러한 게시물 중에는 실제 게시물 내용에 관한 독자의 감상이나 자신의 의견 및 논쟁이 발생하는 댓글도 있는 반면에, 한사람이 지나치게 많은 댓글을 작성하는 경우도 종종 포함되어 있다.

본 논문에서는 게시물의 댓글 집합이 가지는 특성을 분석하고 이를 조회 수에 반영하여 어떠한 상관관계를 가지는지를 분석하고자 한다.

2. 관련 연구

인터넷 게시물의 댓글에 초점을 맞춘 연구로 뉴스 기사에 대한 댓글 토론 구조 추출 연구를 들 수 있다[1]. 5개의 사이트를 중심으로 기사나 블로그의 본문 당 댓글의 개수와 댓글 길이, 댓글 작성자 수에 대해 분석하였다. 하지만 조회 수와 댓글 개수와의 상관관계에 관해 결론짓기에는 조회 수에 관한 언급이 없고, 블로그나 기사의 특성상 댓글의 하위 댓글 구조가 잘 나타나지 않기 때문에 댓글 트리를 생성하는데 적합하지 못하다. 사회 역학 모델을 이용한 뉴스 인기 예측 연구에서는 SNS의 하나인 Digg라는 포털 사이트를 이용하여 흥미도가 찬반 투표 방식에 영향을 미친다는 것을 밝혀, 어떤 뉴스가 더 많은 찬성표를 받을 것인지를 예측하였다[2].

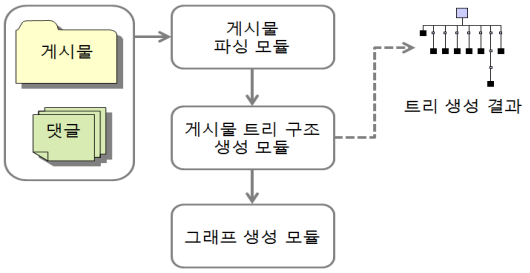
사용자 기반 온라인 기사 인기도 예측 연구에서는 온라인 기사의 댓글이 특정 요일이나 시간에 따라 집중되어 나타나는 점을 착안하여 연구를 진행하였으며, 선형적인 형태를 나타낸다는 것을 밝혔다[3]. 이와 유사하게 온라인 게시 글의 콘텐츠 특성과 조회 수 사이의 관계 연구에서 게시물 콘텐츠의 제목이나 암시적인 내용을 통해 조회 수가 올라가는 것을 보였다[4].

인터넷 커뮤니티의 게시물이나 블로그에 관한 연구에 비해 댓글의 분석 및 댓글 트리구조 시각화에 대한 연구는 상대적으로 그 수가 적으며, 댓글과 조회 수와의 상관관계에 대한 분석 및 연구가 부족한 실정이다. 블로그 탐색 및 시각화 연구[5]에서 제안한 블로그 시각화 도구는 게시물과 댓글의 현황은 알 수 있지만, 게시물과의 관계를 파악할 수 없으므로 적합하지 못하고, 인터넷 게시물의 댓글 분석 및 시각화[6]는 댓글의 일부 단어를 이용하여 필터링한 것을 시각화했기 때문에 댓글의 확장을 시각화 하는 점에서는 적합하지 못하다. 본 연구에서는 시간과 댓글의 상하관계를 이용한 댓글 트리 시각화를 이용하여, 조회 수와의 상관관계를 분석한다.

3. 연구 계획 수립

온라인상의 게시물 가운데 조회 수와 댓글의 확장 사이의 관계는 개인 블로그 보다 게시판이 좀 더 활성화 되어 있는 점을 착안하여, 'PGR21'의 자유게시판 게시물을 수집한다. 게시물에 달린 댓글들을 이용하여 필요한 정보를 수

집하고 게시물마다 댓글 트리를 시각화 하고, 여러 개의 게시물에 포함된 특성을 찾아내어 이를 그래프로 시각화 하여 상관관계를 분석하였다. 연구 계획은 그림 1과 같다.



(그림 1) 상관관계 분석을 위한 연구 계획

연구를 진행하기 위한 과정으로 게시물 파싱, 게시물 트리 구조 생성, 그래프 생성 순으로 진행하였다. 게시물 파싱 과정에서는 자유 게시판에 등재된 게시물 중 약 2주 정도의 게시물들을 일괄적으로 수집하였다. 게시물 당 댓글들의 정보를 이용하여 하나의 게시물의 정보 집합인 *DocTreeData*을 생성하고 파일로 저장하였다. 게시물 트리 구조 생성 과정에서는 *DocTreeData*의 정보를 이용하여 게시물 내부에 있는 댓글 트리를 완성하여 시각화한다. 댓글 트리를 생성하는 과정에서 트리 레벨 평균과 분산, External Node, 최대 트리 깊이 등을 추출하여 그래프 정보 집합인 *GraphResult*를 생성한다. 그래프 생성 모듈에서는 두 개의 모듈을 거치면서 정리되어진 일정 이상의 게시물들의 정보를 종합하여, XY 그래프를 그린다. X, Y 축 상관관계를 이용하여 함수적인 상관관계를 확인한다.

4. 데이터 수집

인터넷을 이용하는 사람들의 자발적인 의사표현방식이 인터넷 공간에서 얼마나 많은 댓글들을 생성하는지를 살펴보기 위해 'PGR21' 커뮤니티 자유게시판의 게시물들을 수집, 조사하였다. 자유게시판에는 정치, 사회, 문화 등 다양한 분야의 게시물들이 등록되는 게시판으로 사용자들이 자유롭게 게시물을 작성하거나, 열람할 수 있다.

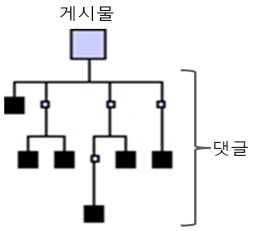
표 1은 자유게시판에 등록된 게시물들의 현황을 나타낸 것이다. 2011년 10월 1일에서 2011년 12월 31일까지 자유게시판에 등록된 게시물을 조사한 것으로 3개월 동안 총 게시물수는 1,972개로 하루 평균 약 21개의 게시물이 등록되었다. 그리고 게시물 당 평균 조회 수는 약 3,564건, 평균 댓글 수는 43개, 그리고 조회 수에 대한 댓글 비는 약 1.2%로 나타났다.

5. 게시물 댓글 트리 시각화

인터넷 게시물 당 댓글 트리구조를 시각화하기 위해서 Orthogonal Tree 구조를 이용하였다. 게시물을 댓글 트리의 최상위에 두고, 댓글 상하관계의 유무에 따라 댓글 트리를 구성하였다.

<표 1> 2011년 4/4분기 'PGR21' 자유게시판 게시물 현황

날짜	게시물	조회 수	댓글수	비율
11.10.01~11.10.31	564	2,160,089	27,919	1.29%
11.11.01~11.11.30	700	2,438,827	28,698	1.17%
11.12.01~11.12.31	708	2,429,655	30,105	1.23%
합계	1,972	7,028,571	86,722	1.23%
게시물 당 평균		3,564	43	1.20%



(그림 2) 게시물 및 댓글의 배치

그림 2는 게시물 파싱 과정에서 정보를 획득한 게시물 중 하나에 대한 댓글트리를 시각화한 결과이다. 그림은 1개의 게시물에 총 10개의 댓글이 배치된 화면으로. 해당 댓글 트리에서는 색상 및 크기가 다른 사각형을 이용하여 댓글의 상하관계를 나타낸다. 색상과 크기가 다른 시각화 요소들은 표 2에 나타나 있다.

<표 2> 시각화 표현 요소

시각화 요소	표현 대상
	게시물
	하위 댓글이 있는 댓글
	하위 댓글이 없는 댓글

데이터 수집을 하는 과정에서 저장되는 게시물의 정보인 *DocTreeData*를 이용하여 트리의 상하관계를 구성하고, 생성된 트리의 External Node 깊이를 이용하여 평균과 분산, 댓글을 작성한 사용자의 수, 댓글 트리의 최대 깊이, 조회 수, 게시물 작성시간과 마지막 댓글과의 시간 간격, 그리고 첫 번째 트리 레벨의 노드 수에 대한 정보를 수집하여 조회 수가 댓글의 확장에 미치는 영향을 분석하는데 이용할 수 있도록 저장하는 작업을 내부적으로 수행한다. 시각화된 트리가 간결하게 표시될 수 있도록 모듈에서 구해지는 데이터들은 아래에 출력하도록 하였다.

6. 조회 수 상관관계 그래프 시각화

게시물을 트리로 구성하는 과정에서 얻어지는 정보들을 취합하여 조회 수와의 상관관계를 파악하기 위해, 점그래프 구조를 이용하였다.

그림 3는 생성된 게시물 댓글 트리정보를 취합하여, 조

회 수와 연관성을 보여주는 그래프이다. 조회 수가 증가할수록 댓글 트리 성분의 증감을 분석하기 위해 댓글 트리의 성분을 X축으로, Y축을 조회 수로 하여 상관관계를 분석하였다. 또한, 결과 도출을 위해 추세 선을 그려 연관성이 있는지 파악해 보았다.

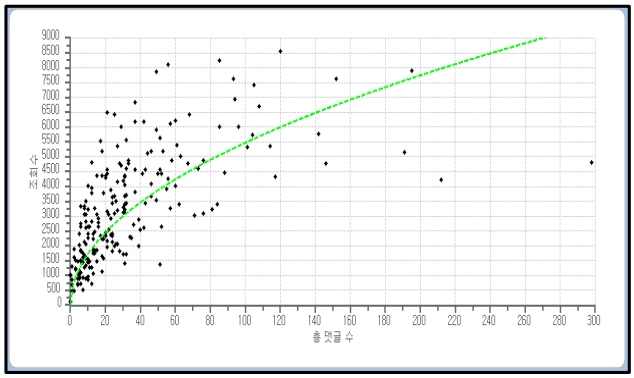
7. 실험 및 결과

게시물 댓글 트리 시각화와 조회수 댓글 트리 성분과의 상관관계 그래프 시각화를 위해, Java와 C#을 이용하였으며, 그래프 라이브러리인 yFiles와 ComponentOne을 사용하였다.

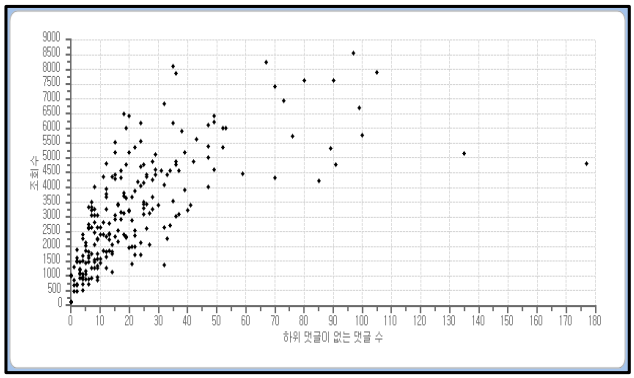
본 논문에서 제안하는 조회 수와 댓글트리의 상관관계를 분석하기 위해 'PGR21'의 자유게시판의 등재된 게시물 중 2주 동안 작성되어있는 자료를 이용하여 게시물을 파싱하고, 게시물 댓글 트리를 구성하였다. 통계적으로 조회 수가 높은 게시물일수록 댓글 트리 성분의 정량적인 수치가 크게 나타나는 점에 착안하여 식을 가정하였다.

$$f(x) = \sqrt{C \bullet x} \qquad \text{(식 1)}$$

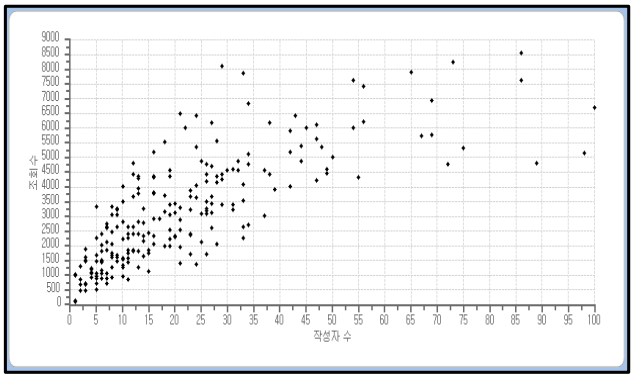
식 1은 제곱근 함수로, 여기서 C는 상수이다. 댓글 트리의 단일 정보의 값이 증가함에 따라 조회 수가 증가하지만, 제곱근 함수의 특징처럼 일정 값 이상으로 댓글 트리 정보의 값이 증가하게 되면 조회 수가 크게 증가하지 않을 것이라고 가정하였다.



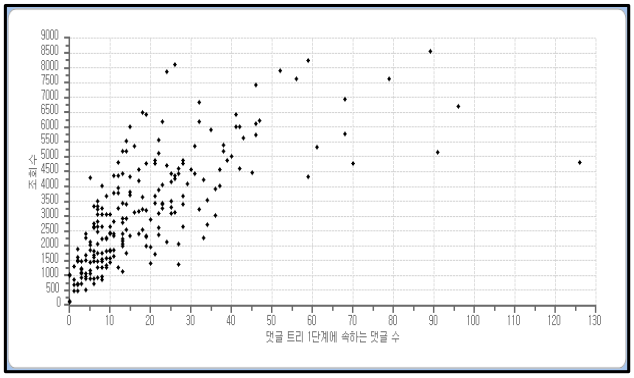
(그림 3) 조회 수와 총 댓글 수와의 상관관계 그래프



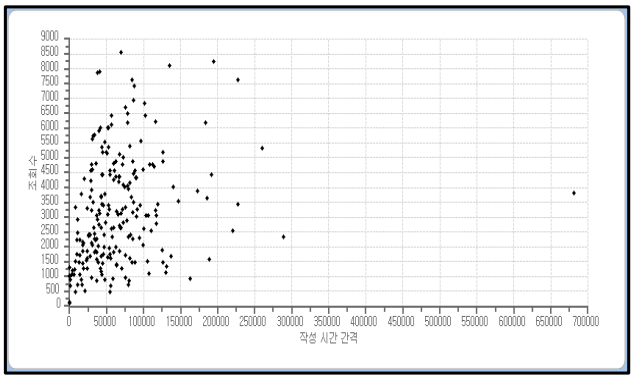
(그림 4) 조회 수와 하위 댓글이 없는 댓글 수 사이의 상관관계 그래프



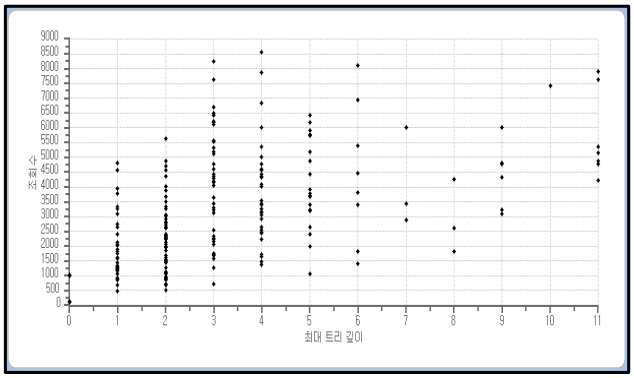
(그림 5) 조회 수와 게시물에 중복되지 않은 작성자 수와의 상관관계 그래프



(그림 6) 조회 수와 게시물 바로 밑에 달린 댓글 수 사이의 상관관계 그래프



(그림 7) 조회 수와 게시물 작성시간과 최종 댓글 작성시간 간격 사이의 상관관계 그래프



(그림 8) 조회 수와 댓글 트리의 최대 깊이와의 그래프

그림 3은 댓글 트리의 정보 중에서 조회 수와 댓글 수가 나타내는 점그래프를 시각화한 것이다. 이때 조회 수가 증가함에 따라 댓글 수가 증가하지만, 추세 선을 그렸을 때 일정한 증가량을 보이지 않는다. 그림 4는 조회 수와 댓글 트리에서 더 이상 댓글이 달리지 않는 댓글들의 숫자를 점그래프로 나타낸 것이다. 그림 5는 댓글을 작성한 작성자의 수와 조회 수를 점그래프로 나타낸 것이다. 그림 6은 댓글 트리에서 깊이가 1인 댓글의 수가 클수록 조회 수가 변화하는 것을 나타낸 점그래프이다. 그림 3, 4, 5, 6은 증가량은 다르지만 각각의 값이 증가할수록 조회 수가 큰 것을 알 수 있다.

그림 7은 게시물 작성 시간과 최종으로 댓글이 작성된 시간과의 시간간격이 조회 수에 미치는 영향을 나타내기 위한 그래프로, 댓글이 달리는 시간과 조회 수는 미치는 영향이 거의 없는 것을 그림에서 확인 할 수 있다.

그림 8은 게시물을 이용하여 구성된 댓글 트리의 최대 깊이의 값이 조회 수에 어떻게 반영되는지 나타낸 그래프이다. 댓글 트리 깊이 값이 크다고 해서 조회 수가 크게 나타나지 않았다.

위의 결과를 통해 인터넷 게시물의 조회 수와 댓글 트리의 정보 중 게시물의 총 댓글수와 하위 댓글이 없는 댓글의 수, 댓글 작성자 수, 댓글 트리 1단계에 위치한 댓글 수 등은 식 1과 같이 조회 수가 증가함에 따라 일정한 수식에 따라 댓글 트리의 성분이 함수적으로 증가하지 않고, 댓글 트리 성분 값이 증가하지만 군집을 이루는 것을 통해 구체화된 상관관계를 나타내기에는 부족하다.

그림 8을 통해 조회 수는 댓글의 수에 대해서는 정량적으로 증가하지만, 댓글의 진행방식에 대해서는 어떠한 영향도 주지 못하는 것을 확인하였다.

조회 수가 유사한 게시물 가운데 댓글 트리 성분의 차이가 큰 경우에 근래에 화제가 되고 있는 사회적 이슈에 관해 댓글 트리 성분 값이 크게 나타나는 경향을 보였다.

8. 결론 및 향후 방향

본 논문에서는 인터넷 게시판을 이용하여 댓글 트리를 시각화하고, 상관관계 그래프를 시각화 및 분석하였다. 이를 기반으로 댓글 트리의 성분에 조회 수가 어떠한 상관관계를 나타내는지 살펴보았다.

분석 결과에 의하면 인터넷 게시물의 조회 수가 증가함에 따라 댓글 트리의 성분 값이 증가하였다. 하지만 특정 함수처럼 일정하게 증가하는 것이 아닌, 인터넷 사용자들의 관심 정도 등 복합적인 요인과 관련되어 증가 폭에 영향을 미치는 것으로 보인다.

본 논문에서 자유게시판의 게시물을 이용하여 댓글 트리를 분석함으로서 게시 글 분류에 따른 분석을 시행하지 않았기 때문에 향후 연구로서 게시 글의 분류에 따라 게시물 댓글 트리의 형태에 대한 시각화를 생각해볼 수 있다. 현재 사회적으로 이슈가 되고 있는 사건에 대해 글이 올라오게 되면 분야에 관계없이 논쟁이 벌어지므로, 게시물에서 논쟁이 진행됨에 따라 댓글 트리의 어떠한 방식으로 벌어지는지 그리고 댓글의 밀도에 대한 분석도 향후 연구로 생각해 볼 수 있다.

참고문헌

[1] Anne Schuth, Maarten Marx, Maarten de Rijke, “Extracting the Discussion Structure in Comments on News-Articles,“ 2007 ACM 978-1-59593-829-9/07/0011, pp.97-104, 2007

[2] K. Lerman and T. Hogg, “Using a model of social dynamics to predict popularity of news,” in Proceedings of the 19th international conference on World wide web. ACM, 2010, pp. 621 - 30.

[3] A. Tatar, P. Antoniadis, M. D. Amorim, “Predicting the popularity of online articles based on user,” 2011 ACM 978-1-4503-0148-0/11/05, 2011

[4] 이원준, “온라인 게시글의 콘텐츠 특성과 조회 수간의 관계”, 한국콘텐츠학회논문지 ′10 Vol. 10 No. 2, pp.241-249, 2010

[5] 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.

[6] 이윤정, 지정훈, 우균, 조환규 “인터넷 게시물의 댓글 분석 및 시각화”, 한국콘텐츠학회논문지 ′09 Vol. 9 No. 7, pp.45-56, 2009