http://dx.doi.org/10.5392/JKCA.2012.12.02.040

인터넷 토론 사이트의 논쟁댓글 및 논쟁관계 시각화

Extracting and Visualizing Dispute comments and Relations on Internet Forum Site

이윤정*, 정인준**, 우 균**
부산대학교 U-Port 정보기술사업단*, 부산대학교 컴퓨터공학과**

Yun-Jung Lee(leeyj01@pusan.ac.kr)*, In-Joon Jung(spd1335@pusan.ac.kr)**,
Gyun Woo(woogyun@pusan.ac.kr)**

요약

최근에는 인터넷 토론 사이트에서 댓글을 이용해 다른 사람들과 토론이나 논쟁하는 경우를 흔히 볼 수 있다. 논쟁을 통해 게시물의 내용과는 다른 새로운 의견이 나타날 수도 있으므로 논쟁댓글을 파악하고 식별하는 것은 중요한 문제라고 할 수 있다. 본 논문에서는 국내의 인터넷 토론 사이트인 SkepticalLeft와 아고라에서 수집한 댓글을 통해 인터넷 토론 게시판에서 논쟁댓글의 특성을 분석하였다. 그리고 이를 바탕으로 댓글 목록의 논쟁구간과 논쟁관계를 검출하고 이를 시각화하는 방법을 제안한다. 제안 방법의 성능을 보이기 위해 논쟁댓글과 논쟁 쌍을 검출하고 정확도와 재현율 그리고 F-measure를 측정하였다. 논쟁댓글 검출 성능은 F-measure가 0.84(SkepticalLeft)와 0.83(아고라)으로 측정되었고, 논쟁 쌍 검출은 각각 0.75(SketpcialLeft)와 0.82(아고라)로 측정되었다. 제안 방법은 댓글 작성자의 순서관계만을 이용하므로 사용언어나 철자법에 제약받지 않는다. 또한 시각화된 뷰를 통해 게시판 이용자들이 댓글에 내포된 논쟁 구조를 파악하는데 도움을 줄 것이다.

■ 중심어 : | 논쟁 댓글 | 온라인 토론 | 논쟁 시각화 | 논쟁 검출 |

Abstract

Recently, many users discuss and argue with others using replying comments. This implies that a series of comments can be a new source of information since various opinions can be appeared in the dispute. It is important to understand the implicit dispute structure immanent in the comment set. In this paper, we examine the characteristics of disputes using replying comments in the Internet forum sites using a set of test articles with the comments collected from SketicalLeft and Agora, which are famous Internet forum sites in Korea. And we propose a new method for detecting and visualizing the dispute sections and relations from a large set of replying comments. To show the performance of our method, we measured precision, recall, and F-measure. According to the experimental results, the F-measures of the detection of the comments in dispute are about 0.84 (SketpcialLeft) and 0.83 (Agora); those of the detection of the commenter pairs in dispute are 0.75 (SketpcialLeft) and 0.82 (Agora), respectively. Since our method exploits the temporal order of commenters to detect the disputes, it is not dependent on the host language nor on the typos in comments. Also, our method can help the readers to grasp the structure of controversy hidden in the comment set through the visualized view.

■ keyword : | Dispute Comment | Online Dispute | Dispute Visualization | Dispute Extraction |

* 이 논문은 2011년도 정부재원(교육과학기술부 인문사회연구역량강화사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2011-371-B00008)

접수번호 : #111213-009
접수일자 : 2011년 12월 13일
심사완료일 : 2012년 01월 19일
교신저자 : 우 균, e-mail : woogyun@pusan.ac.kr

Ⅰ. 서 론

온라인 커뮤니티의 사용이 증가함에 따라 온라인상으로 서로의 의견을 교환하거나 토론하는 경우가 많아지고 있다. 온라인 토론은 크게 메신저와 같은 일시적인 쌍방향 메시지 전달 형태와 인터넷 토론 사이트와 같은 공개 토론의 형태로 나누어 볼 수 있다. 특히 인터넷 토론 사이트의 경우 특정 주제에 대한 여러 사람들의 의견을 살펴보거나 게시물 등록이나 댓글 달기와 같은 형태로 의견을 공유할 수 있어 새로운 토론 매체로 활용되고 있다.

온라인 토론에서 댓글은 자신이 읽은 게시물에 대한 가장 적극적인 의사 표현 형태로 볼 수 있으며, 게시물과 이용자들과 상호작용할 수 있는 쉽고 효율적인 방법으로 사용되고 있다[1]. 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[2]. 실제로 많은 인터넷 토론 사이트에서 짧은 댓글을 이용해 실시간 논쟁이 일어나는 경우를 흔히 볼 수 있다[3][4].

사이트마다 차이는 있으나 대부분의 경우 댓글은 길이가 짧으며 게시물에 대한 독자의 감상을 나타내는 것이 많지만 때로는 댓글을 통해 게시물과는 다른 자신의 의견을 나타내기도 하며 다른 이용자들과 논쟁을 벌이기도 한다. 이용자들은 댓글을 읽음으로써 게시물에 대한 독자들의 전반적인 의견을 살펴볼 수도 있으며 게시물에 나타나지 않은 새로운 정보를 얻을 수도 있다. 따라서 댓글은 게시물과는 다른 새로운 정보원으로 사용될 수 있다.

인터넷 게시물의 댓글에 대한 기존 연구로 Mishne과 Glance의 연구를 들 수 있다[5]. 이 연구에서는 댓글은 블로그 공간에서 게시물 양의 약 30% 정도로 상당한 비중을 차지하고 있으며, 댓글의 양은 블로그나 게시물에 대한 관심 정도를 가리키는 지시자로 사용될 수 있음을 제안하였다. 또한 댓글의 내용을 분석하여 논쟁 정도를 계산하는 방법을 제시하였다. 댓글의 통계적 특성에 관한 연구도 제시되었다[6-8]. 이윤정 등은 대규모 댓글의 시각화를 위한 TRIB 시스템을 제안하였다. TRIB는 인기 게시물의 경우 수 천 개 이상의 댓글이 달리는 등 지금의 게시물과 게시물에 달린 댓글의 의미적 관계를 고려하여 전체 댓글을 하나의 뷰로 시각화한다. 하나의 게시물에 많은 댓글이 달린 경우에도 이용자들의 쉽게 게시물과 댓글과의 관계를 파악할 수 있다는 장점이 있다[6][7]. 그들의 다른 연구에서는 댓글 작성자 별 댓글 수의 분포가 거듭제곱 법칙을 따른다는 통계적 특성을 이용하여 스킵 리스트를 기반으로 하는 댓글 관리 방법을 제안하였다[8]. 실험을 통해 AVL 트리보다 스킵리스트가 댓글의 검색, 삽입, 삭제 성능이 더 나음을 보였다.

최근에 사회 연결망 서비스나 블로그에 관한 연구 중에 게시물이나 댓글의 내용을 요약하거나 의견을 추출하는 연구가 보고되었다. Hu 등은 블로그에서 댓글에서 토의된 주제에 대한 대표 문장을 추출하는 방법을 제안하였다[9]. 그리고 Zhou와 Hovy는 온라인 토론과 블로그에서 소개된 정보를 동적으로 요약하는 방법을 제안하였다[10]. 그러나 이러한 연구들은 댓글의 내용 분석을 기반으로 하고 있어 사용 언어에 제한적인 단점이 있다.

본 논문에서는 댓글에서 나타나는 여러 현상 중 논쟁에 대해 주안점을 둔다. 논쟁을 통해 게시물의 내용과는 다른 새로운 의견이 나타날 수도 있으므로 논쟁댓글을 파악하고 식별하는 것은 중요한 문제라고 할 수 있다. 그러나 모든 댓글을 읽어보고 논쟁의 유무나 논쟁관계를 파악하는 것은 상당한 노력이 필요하다. 전체 댓글 목록에서 논쟁이 벌어지는 구간이 어디인지, 논쟁 정도는 어떠한지, 그리고 어떤 작성자들이 논쟁을 벌이는지에 대한 정보를 이용자들이 댓글을 읽기 전에 사전 정보로 제공하고자 한다.

이를 위해 본 논문에서는 댓글 집합에 포함된 논쟁구간을 식별하고 논쟁관계를 추출하는 방법을 제안한다. 제안 방법은 댓글에서 논쟁이 일어날 경우 논쟁중인 작성자들의 글이 번갈아 나타날 확률이 높다는 점에 착안하였다. 따라서 댓글의 내용 분석을 기반으로 하는 기존의 방법들과는 달리, 댓글 작성순서와 댓글 작성자

정보만을 이용하여 댓글의 논쟁을 식별하므로 이 방법은 댓글 작성 언어와 무관하게 사용할 수 있으며 철자법 오류와 같은 문제에 구애받지 않는다. 또한 검출한 작성자 간의 논쟁관계를 비방향성 그래프로 시각화함으로써 이용자들이 댓글을 읽어보기 전에 전체적인 논쟁 정도와 논쟁관계를 직관적으로 파악할 수 있도록 해준다.

본 논문의 구성은 다음과 같다. 2장에서는 논쟁댓글의 특성에 대해서 설명한다. 3장에서는 댓글에 포함된 논쟁구간 및 논쟁관계 검출에 대해 자세히 설명한다. 4장에서는 논쟁관계 시각화에 대해 설명하고 5장에서는 실험 결과를 보이고 6장에서 결론을 맺는다.

II. 논쟁댓글의 특성

제안 시스템의 기본 아이디어는 논쟁댓글의 특성을 관찰을 기반으로 한다. 일반적으로 한 게시물에 달린 일련의 댓글에서 논쟁을 벌이고 있는 구간은 특정 작성자들의 글이 번갈아 나타나는 경향이 있다. 그들은 댓글을 통해 각자의 의견을 나타내거나 상대방의 의견에 대해 반박하며 논쟁을 이어간다. [그림 1]은 인터넷 토론 게시판에서 흔히 볼 수 있는 댓글에서의 논쟁이다. [그림 1]은 인터넷 토론 게시판인 SkepticalLeft 사이트에서 스크랩한 것으로 두 이용자가 댓글로 서로 논쟁중인 것을 보여준다[12].

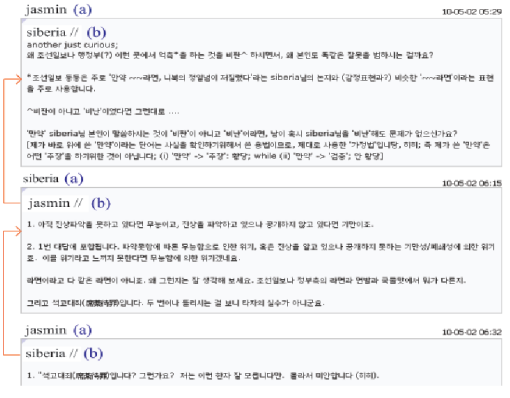


그림 1. 인터넷 토론 게시판에서 댓글을 이용한 논쟁

온라인 토론 사이트에서 댓글을 통한 논쟁은 메신저와 같은 채팅 서비스처럼 실시간에 이루어지는 것은 아니지만 논쟁중인 작성자의 글이 논쟁구간에서 자주 반복적으로 나타날 확률이 높으며, 논쟁 중인 작성자들은 그렇지 않을 때보다 댓글 작성횟수가 많을 것으로 예상된다.

본 논문에서는 실제 인터넷 토론 사이트의 게시물에 달린 댓글 분석을 통해 위와 같은 가정이 타당한지를 살펴보았다. 먼저 얼마나 많은 댓글과 댓글 작성자들이 논쟁에 참여하였는지를 살펴보기 위해 SkepticalLeft 사이트에서 게시물과 게시물에 달린 댓글을 수집하고, 논쟁에 참여한 댓글과 그렇지않은 것으로 분류하였다. [표 1]은 20개의 게시물에 달린 댓글집합을 간략히 정리한 것이다.

표 1. 'SkepticalLeft' 게시물의 댓글 통계.

	일반	논쟁	합계
댓글	152	2,420	2,572
댓글 작성자	93	233	326
1인당 평균 댓글 작성 횟수	1.6	11.3	8.2

총 댓글 수는 2,572개로 그 중에서 논쟁에 참여한 댓글은 전체의 약 94%인 2,420개이다. 1인 당 평균 댓글 작성 횟수는 논쟁에 참여한 작성자들이 11.3회로 그렇지 않은 작성자들보다 훨씬 높게 나타난 것으로 보아 논쟁중인 작성자들의 댓글참여 빈도가 높을 것이라는 우리의 가정이 타당함을 알 수 있다.

다음으로 논쟁중인 작성자의 댓글이 논쟁구간에서 얼마나 가깝게 나타나는지를 알아보기 위해 논쟁에 참여한 작성자들의 댓글 거리를 조사하였다. 본 논문에서는 한 게시물에 달린 댓글 집합을 $C=\{c_i | i \in N\}$로 정의한다. 여기서 N은 해당 게시물에 달린 전체 댓글 수를 나타내고, c_i는 등록된 시간상으로 i번째에 해당하는 댓글을 가리킨다. 다음으로 $w(c_i)$는 댓글 c_i를 쓴 작성자를 나타낸다. 마지막으로 $cdist(c_i)$는 댓글 거리를 나타내는 것으로 c_i와 $w(c_i)$가 c_i 전에 쓴 댓글과의 최소 거리를 의미하며 식 1과 같이 구할 수 있다.

$$cdist(c_i) = \begin{cases} \infty & \text{if } i = 1 \\ \min\{d(c_j, c_i) | c_j \in C, 1 \le j < i\} \end{cases}$$
$$d(c_j, c_i) = \begin{cases} i - j & \text{if } w(c_j) = w(c_i) \\ \infty & \text{if } w(c_j) \neq w(c_i) \end{cases} \quad (1)$$

이때 동일한 작성자가 쓴 연속된 댓글은 댓글의 하나의 댓글로 간주한다. 왜냐하면 일반적으로 인터넷 토론 사이트에서는 댓글의 글자 수를 300자 이내 등으로 제한하는 경우가 많아 댓글의 내용이 길어질 경우 다음 댓글로 넘겨 계속 적는 경우가 많기 때문이다.

[그림 2]는 [표 1]의 데이터 집합에서 논쟁에 참여한 작성자들이 쓴 댓글을 대상으로 계산한 댓글거리 $cdist$의 분포를 보여준다.

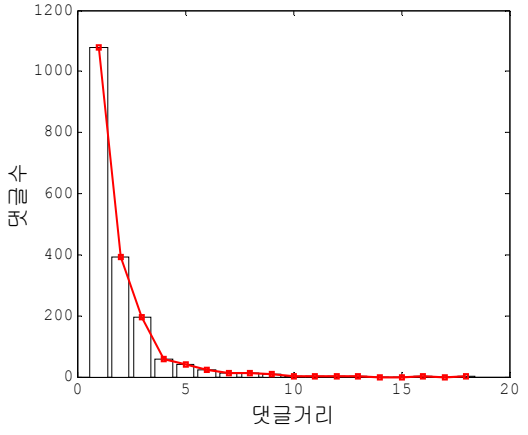


그림 2. 표 1의 게시물에서 논쟁에 참여한 작성자들이 쓴 댓글의 댓글거리 분포

[그림 2]에서 확인할 수 있듯이 대부분 댓글들의 댓글 거리는 5이하로 나타났다. 이것은 한 사람이 동일한 게시물에 여러 번 댓글을 다는 경우 이전에 쓴 댓글과 인접해서 다시 쓸 확률이 높다는 것을 의미한다. [표 2]는 댓글거리의 통계 정보를 보여준다.

표 2. 표 1의 게시물에서 논쟁에 참여한 작성자들이 쓴 댓글의 댓글거리 통계

	평균	분산	최대	최소
댓글거리	1.87	2.76	18	1

Ⅲ. 논쟁 검출

댓글 내용을 읽지 않고 댓글 목록에 포함된 논쟁구간이나 논쟁관계를 추출하기 위해서는 각 댓글이 지닌 논쟁의 가능성을 정규화 된 수치로 표현하는 방법이 필요하다. 본 논문에서는 댓글 작성자들의 순서 관계를 이용하여 댓글의 논쟁 가능성을 논쟁 지수로 표현하고 이 값을 이용하여 전체 댓글에서의 논쟁구간과 논쟁중인 작성자 쌍을 검출하는 방법을 제안한다.

1. 논쟁구간 검출

댓글 집합에서 논쟁구간을 검출하기 위해서 먼저 댓글의 논쟁 가능성을 구하고 해당 댓글이 논쟁댓글인지를 판별해야 한다. 앞서 [그림 2]에서 확인한 댓글거리 분포로 보아 댓글거리가 짧은 댓글일수록 논쟁댓글일 확률이 높다고 할 수 있다. [그림 3]은 논쟁거리 분포를 지수함수로 추정한 확률분포의 그래프를 보여준다.

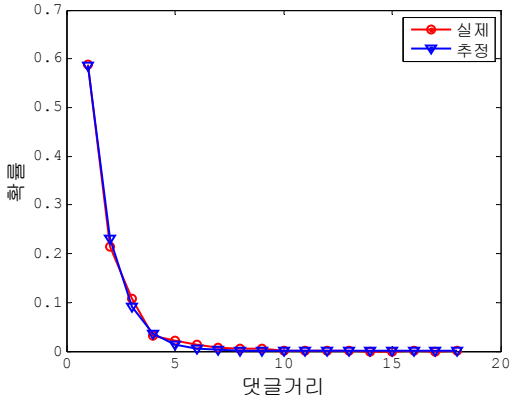


그림 3. 실제 논쟁댓글의 댓글거리 분포와 지수함수로 추정한 확률분포

그림을 통해 댓글의 실제 논쟁거리의 분포가 지수함수로 잘 적합화 됨을 알 수 있다.

본 논문에서는 댓글의 논쟁 확률을 논쟁지수라고 하고 $DI(c_i)$와 같이 표기한다. 논쟁지수는 식 2와 같이 구할 수 있다.

$$DI(c_i) = \begin{cases} \alpha \bullet e^{\beta(cdist(c_i)-1)}, & \text{if } DI(c_i) \geq \theta \\ \mu & , \ otherwise \end{cases} \quad (2)$$

여기서 α와 β는 논쟁 확률 조정 매개변수로서 실험을 통해 구할 수 있다. 식 2에 따르면 댓글거리가 멀어질수록 논쟁지수 값이 0에 가까워지는데 이때 θ를 임계치로 하여 계산된 논쟁지수가 θ이하이면 음수인 μ 값으로 대체한다. 그 이유는 제안방법에서 논쟁구간 식별을 위해 논쟁지수의 합을 구하는데 이때 의미 없는 0에 가까운 논쟁지수가 계속 더해지는 것을 방지하기 위해서이다.

제안 방법에서는 전체 댓글에서 각 댓글의 논쟁지수의 합이 가장 큰 구간을 논쟁구간으로 간주한다. 댓글 c_i까지 누적논쟁지수를 $CDI(c_i)$라고 하고 식 3과 같이 구할 수 있다.

$$CDI(c_i) = \max_{k=1}^{i-1} \left\{ \sum_{j=k}^{i} DI(c_j),\ 0 \right\} \quad (3)$$

일단 최대 논쟁구간이 구해지면 해당 구간에 포함되는 댓글의 논쟁지수를 최솟값 μ로 바꾸어 다시 CDI가 최대가 되는 구간을 찾으면 그 다음으로 논쟁확률이 높은 구간을 찾을 수 있게 된다. 이런 과정을 반복적으로 수행하면 논쟁 확률이 높은 구간을 순서대로 찾을 수 있다. 이해를 돕기 위해 [그림 4]와 같은 28개의 댓글 시퀀스를 가정하자.

A B A B C D C E C F G H G H G H G H I G C I C I A C J K

그림 4. 댓글 등록 시간의 오름차순으로 정렬된 댓글 시퀀스 예. 각각의 알파벳은 댓글 작성자 ID를 나타낸다.

그림에서 각각의 알파벳은 댓글 작성자의 ID를 나타내고 댓글은 등록된 시간의 오름차순으로 정렬되어 있다. 이 댓글 시퀀스에서 댓글들의 논쟁지수 $DI(c_i)$와 누적논쟁지수 $CDI(c_i)$는 [표 3]과 같다. 이때 논쟁지수 $DI(c_i)$를 계산하기 위해서 식 2의 제어 매개변수 α, β와 최소 논쟁지수 μ, 유효논쟁지수를 위한 임계치 θ는 SkepticalLeft의 댓글 집합에서 실험을 통해 얻어진 0.7084와 -0.5655, 그리고 -0.5와 0.001로 각각 정의하였다. 그림 5는 표 3의 댓글 시퀀스에 대한 논쟁지수와 누적논쟁지수의 그래프를 보여준다. [그림 5]에서 가로축은 댓글 순서를 의미하며, 세로축은 해당 댓글의 논쟁지수와 누적논쟁지수 값을 가리킨다.

표 3. 그림 4에 나타난 댓글 시퀀스의 논쟁지수

i	$w(c_i)$	$cdist(c_i)$	$DI(c_i)$	$CDI(c_i)$
1	A	∞	-0.500	0.000
2	B	∞	-0.500	0.000
3	A	2	0.402	0.402
4	B	2	0.402	0.804
5	C	∞	-0.500	0.304
6	D	∞	-0.500	0.000
7	C	2	0.402	0.402
8	E	∞	-0.500	0.000
9	C	2	0.402	0.402
10	F	∞	-0.500	0.000
11	G	∞	-0.500	0.000
12	H	∞	-0.500	0.000
13	G	2	0.402	0.402
14	H	2	0.402	0.804
15	G	2	0.402	1.206
16	H	2	0.402	1.608
17	G	2	0.402	2.010
18	H	2	0.402	2.412
19	I	∞	-0.500	1.912
20	G	3	0.229	2.141
21	C	12	0.001	2.142
22	I	3	0.229	2.371
23	C	2	0.402	2.773
24	I	2	0.402	3.175
25	A	22	-0.500	2.675
26	C	3	0.229	2.904
27	J	∞	-0.500	2.404
28	K	∞	-0.500	1.904

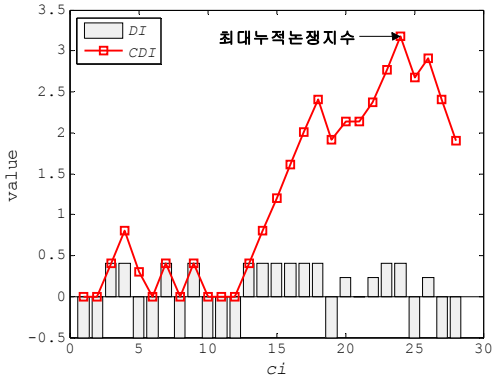


그림 5. 표 3에 나타난 댓글 시퀀스의 논쟁지수와 누적논쟁지수 그래프

그래프에서 최대누적논쟁지수는 c_{24}일 때의 3.175 이다. 이때 최대논쟁구간은 최대CDI를 기록한 댓글에서 댓글 순서상 앞쪽으로 CDI가 0이 되는 지점까지의 구간으로 정한다. 따라서 이 댓글 시퀀스에서는 c_{12}에서 c_{24}까지가 최대 논쟁구간이 된다. [그림 6]은 [그림 4]의 댓글 시퀀스에서 최대논쟁구간으로 간주되는 구간을 보여준다.

..... H G H G H G H I G C I C I .....

그림 6. 그림 4의 댓글 시퀀스에서 최대 논쟁구간

그림의 댓글 시퀀스에서 댓글 작성자 H와 G, I와 C가 서로 논쟁하고 있을 것으로 추정해 볼 수 있다.

2. 논쟁관계 검출

일반적으로 댓글을 이용한 논쟁은 누군가의 댓글에 대한 자신의 의견을 또 다른 댓글로 표현하는 형식을 띠므로 논쟁관계에 있는 작성자들의 댓글은 서로 인접해서 나타나는 경향이 있다. 본 논문에서는 작성자들의 댓글이 얼마나 자주 인접해서 나타나는지를 작성자 결합도라고 정의하고 이것을 이용하여 댓글 작성자들의 논쟁관계를 검출한다. 두 작성자 ID_1와 ID_2의 결합도 $CW(ID_1, ID_2)$는 식 4와 같이 구할 수 있다.

$$CW(ID_1, ID_2) = \gamma \cdot e^{\delta(wdist(ID_1, ID_2))} \cdot f(ID_1, ID_2) \quad (4)$$

여기서 γ와 δ는 작성자의 논쟁 확률 제어 매개변수로서 실험을 통해 구할 수 있다. $wdist(ID_1, ID_2)$는 식 1에서의 $cdist$와 비슷한 개념으로 두 작성자가 쓴 댓글들의 댓글거리를 의미한다. 일반적으로 두 개 이상의 댓글이 번갈아 나타나므로 평균 댓글거리로 정의한다. 예를 들어 [그림 4]의 댓글 시퀀스에서 작성자 A와 C의 댓글거리 $wdist(A, C)$는 [그림 7]에서와 같이 약 1.67이다.

그림 7. 작성자 A와 C의 작성자 댓글거리 $wdist(A, C)$

이때 두 작성자의 댓글이 번갈아 나타나는 것만을 논쟁댓글로 고려하므로 c_1, c_7, c_9 그리고 c_{21}은 댓글거리 계산에서 제외시킨다. 식 4에서 $f(ID_1, ID_2)$는 두 작성자의 글이 번갈아 나타나는 횟수를 나타내므로 $f(A, C)$는 3이 된다.

작성자 논쟁관계를 검출하기 위해서 앞서 찾아진 논쟁구간내의 모든 작성자 쌍을 구하고 각 쌍에 대해 식 4를 이용하여 작성자 결합도를 구한다. [그림 6]의 논쟁구간에는 G, H, I 그리고 C의 작성자가 있으며 모두 6개의 작성자 쌍을 찾을 수 있다. 모든 작성자 쌍의 결합도는 [표 4]와 같다.

표 4. 논쟁구간에 포함된 댓글 작성자 쌍의 결합도

	G	C	H	I
G		0.20	1.32	0.43
C	0.20		0.07	0.72
H	1.32	0.07		0.20
I	0.43	0.72	0.20	

작성자들의 쌍에서 작성자의 순서는 관계없다. 즉 $\langle G, H \rangle$와 $\langle H, G \rangle$는 같은 쌍을 의미한다. 제안 방법에서는 유효한 논쟁 쌍을 검출하기 위해서 임계치 이상의 결합도를 가지는 작성자 쌍을 논쟁관계가 있는 것으로 간주한다. [표 4]에서 임계치를 0.5로 설정한다면 $\langle G, H \rangle$와 $\langle C, I \rangle$ 쌍이 논쟁 쌍으로 검출된다.

IV. 논쟁 시각화

본 논문에서는 댓글 집합에 포함된 논쟁을 직관적으로 파악할 수 있도록 하기 위해서 앞서 검출된 논쟁구간과 논쟁관계를 시각화한다. 논쟁구간은 각 댓글의 누적논쟁지수를 이용하여 [그림 8]과 같이 파형 그래프로 시각화한다.

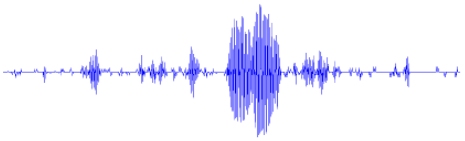


그림 8. 댓글 시퀀스의 누적논쟁지수를 이용한 논쟁구간 시각화

[그림 8]에서 가로축은 댓글의 순서를 가리키고 세로축인 그래프의 진폭은 해당 댓글의 누적논쟁지수를 가리킨다. 따라서 진폭이 큰 구간이 누적논쟁지수가 높은 구간이므로 쉽게 논쟁구간을 식별할 수 있다. [그림 8]에서는 한 구간만 강한 논쟁의 가능성이 있는 구간으로 나타나지만 댓글 집합에서 논쟁이 활발한 경우에는 비슷한 강도의 논쟁이 지속적으로 또는 여러 부분에서 나타날 수도 있다.

본 논문에서는 논쟁구간에서 찾아진 작성자들의 논쟁관계는 비방향성 그래프로 시각화하고 이것을 논쟁관계 그래프라고 정의한다. 논쟁관계 그래프에서 각각의 작성자가 노드로 표현되며 작성자들의 결합도는 그 노드를 잇는 에지의 강도로 표현된다. [그림 9]는 [표 4]에서 나타난 논쟁 쌍들로 구성된 논쟁관계 그래프를 보여준다.

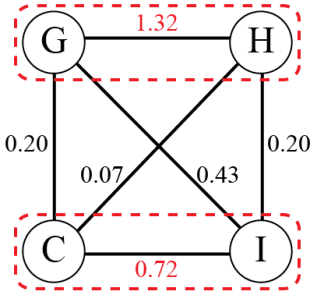


그림 9. 작성자 쌍의 결합도를 이용하여 표현한 비방향성 그래프

[그림 9]에서 보듯이 모든 가능한 논쟁관계는 전체 4개의 노드로 구성된 비방향성인 완전연결 그래프로 나타나고 임계치를 0.5로 설정할 경우 점선으로 표현된 노드 쌍인 〈G,H〉와 〈C,I〉만 남는다. 최종적으로 시각화 된 논쟁관계 그래프는 [그림 10]과 같다.

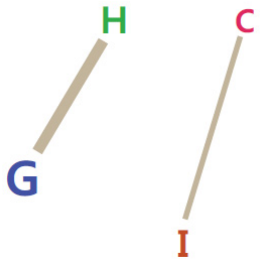


그림 10. 논쟁관계 시각화의 예

[그림 10]에서 노드는 댓글 작성자의 ID를 나타내고 노드의 크기는 해당 작성자가 쓴 댓글의 수에 비례한다. 노드를 연결한 에지의 두께는 두 작성자간의 결합도를 나타낸다. 그래프에서 노드의 색상은 중복되지 않게 무작위로 선택되며, 노드의 위치는 그래프 배치 알고리즘 중의 하나인 Kamada-Kawai 방법으로 정하였다[11]. 따라서 논쟁관계 그래프를 통해서 작성자들의 논쟁관계뿐만 아니라 논쟁의 강도 및 댓글 작성 횟수까지 한눈에 파악할 수 있다.

V. 실험 및 결과

본 논문에서 제안하는 댓글 집합에 포함된 논쟁구간과 논쟁관계 검출의 성능을 보이기 위해서 두 개의 인터넷 토론 게시판 SepticalLeft와 아고라에서 댓글을 수집하였다[12][13]. SkpeticalLeft는 토론을 위주로 하는 사이트로 댓글의 길이가 길고, 댓글 작성자들이 토론에 참여하는 비율이 높으며 댓글을 달 때 자신이 논쟁하고자 하는 댓글 작성자의 ID를 댓글 첫머리에 표시하는 것이 일반적이다. 따라서 댓글의 논쟁 참여 여부와 논쟁관계를 명시적으로 파악할 수 있다.

반면에 아고라 게시판은 300자 이내의 댓글을 허용하고 있어서 댓글의 길이가 짧으며, 논쟁에 참여하는 댓글이라도 누구와 논쟁하는지 명시적으로 밝히고 있지 않은 경우가 대부분이라 댓글 내용으로 논쟁관계를 파악해야 한다. 인기 게시물의 경우 수백 개 이상의 댓글이 달린 경우도 적지 않게 찾아볼 수 있어 논쟁뿐만 아니라 중복 댓글이나 스팸과 같은 댓글에서 일어나는 다

양한 현상을 볼 수 있다.

실험을 위해 두 사이트에서 40개의 게시물과 그 게시물에 달린 댓글을 수집하고, 댓글 내용을 기반으로 논쟁댓글 여부를 판별하였다. [표 5]는 실험 데이터의 분포를 정리한 것이다.

표 5. 실험 데이터 집합에 포함된 논쟁댓글 분포

	SkepticalLeft	아고라
총 댓글 수	2,284 (100%)	3,383 (100%)
논쟁 댓글	1,621 (71%)	1,561 (46%)
일반 댓글	663 (29%)	1,182 (54%)

[표 5]에서 논쟁댓글은 논쟁에 참여한 댓글을 가리키며, 일반 댓글은 논쟁과는 무관한 댓글을 나타낸다. 댓글 내용에 논쟁의 대상이 명시되어 있거나 내용을 읽어 논쟁의 대상을 알 수 있는 경우에 논쟁댓글로 분류하였다. 논쟁 댓글 집합이 정해지면 검출 결과는 [표 6]과 같이 분류할 수 있다.

표 6. 논쟁댓글 검출에서 가능한 분류 결과

		실제 조건	
		논쟁(P)	일반(N)
검출 결과	논쟁(P)	TP	FP
	일반(N)	FN	TN

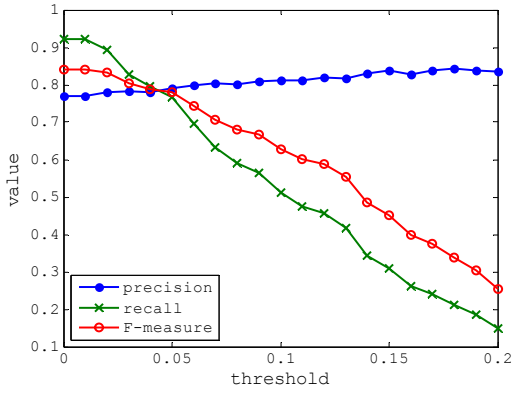
[표 6]에서 T와 F는 각각 참(true)과 거짓(false)을 의미하는데, T는 실제 상태와 검출 결과가 같음을 의미하고 N은 실제 상태와 검출 결과가 다름을 의미한다. P와 N은 각각 양성(positive)과 음성(negative)을 나타내는데, 본 실험에서 P는 논쟁댓글을 N은 일반댓글을 의미한다. 따라서 TP는 실제 논쟁인 댓글을 논쟁 그룹으로 분류한 것을 나타내고, FN은 실제 논쟁댓글을 논쟁이 아닌 것으로 분류한 것을 나타낸다.

본 논문에서는 제안 방법의 성능을 정량적으로 측정하기 위해서 정보검색 분야의 성능 측정 기준으로 주로 사용되는 방법인 정확도(precision)와 재현율(recall)을 사용한다. 정확도와 재현율을 구하는 방법은 아래와 같다.

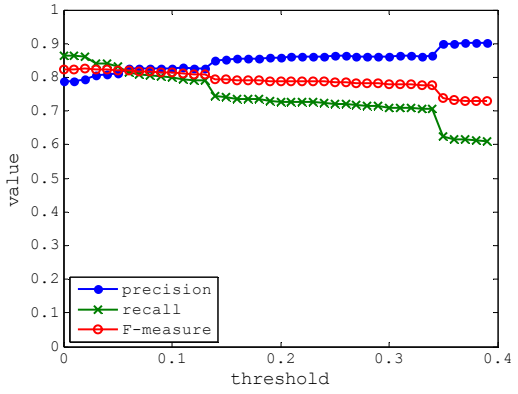
$$preision(\theta) = \frac{TP}{TP+FP}$$

$$recall(\theta) = \frac{TP}{TP+FN}$$

제안 방법의 논쟁댓글 검출 성능을 알아보기 위해 데이터 집합에서 유효논쟁지수의 임계치를 달리하여 논쟁댓글 검출 결과의 정확도와 재현율 그리고 F-measure를 측정하였다. [그림 11]은 제안 방법의 논쟁댓글 검출 결과를 보여준다. [그림 11]에서 가로축은 유효논쟁을 식별하는 임계치 θ를 가리키고, 세로축은 정확도, 재현율 그리고 F-measure를 나타낸다. 두 사이트 모두 임계치가 높을수록 정확도는 증가하고, 재현율은 감소하는 것으로 나타난다. 즉 임계치를 높이면 강한 논쟁 즉 논쟁지수가 높은 것들만 검출됨을 의미한다.



(a) SkepticalLeft



(b) 아고라

그림 11. 유효논쟁지수 임계치 변화에 따른 논쟁댓글 검출 성능

[표 7]은 정확도, 재현율 그리고 F-measure가 모두 높을 때의 값들을 정리한 것이다.

표 7. 제안 방법의 논쟁댓글 검출 성능

site	θ	precision	recall	F-measure
SkepticalLeft	0.01	0.77	0.92	0.84
아고라	0.02	0.79	0.86	0.83

제안 방법의 논쟁댓글 검출 성능은 SkepticalLeft의 경우 논쟁지수의 임계치를 0.01로 했을 때 F-measure가 0.84로 최대로 나타났고, 아고라의 경우는 임계치를 0.02로 했을 때 F-measure가 0.83으로 나타났다.

다음으로 제안 방법의 논쟁 작성자 쌍의 검출 성능을 알아보기 위해 데이터 집합에서 논쟁관계에 있는 모든 작성자 쌍을 수작업으로 판별하였다. [표 8]은 논쟁 작성자 쌍을 정리한 것이다.

표 8. 실험 데이터 집합에 포함된 논쟁 작성자 쌍 분포

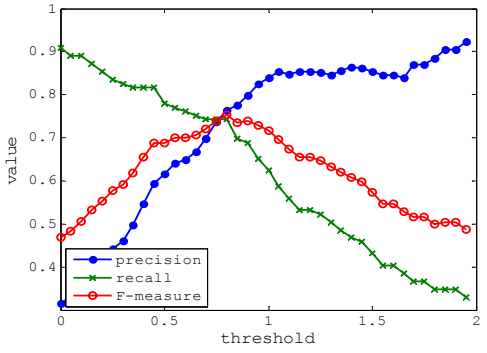
	SkepticalLeft	아고라
총 작성자 쌍	1,535 (100%)	7,603 (100%)
논쟁 작성자 쌍	109 (7.1%)	33 (0.4%)
일반 작성자 쌍	1,426 (92.9%)	7,570 (99.6%)

SkepticalLeft의 경우 전체 작성자 쌍 중에서 약 7.1%인 109개 쌍이 논쟁관계가 있으며, 아고라의 경우는 전체의 약 0.4%인 33개 쌍이 논쟁관계가 있는 것으로 조사되었다. 논쟁댓글 검출 실험과 마찬가지로 유효논쟁지수의 임계치를 변화시키며 논쟁 작성자 쌍 검출의 정확도와 재현율을 측정하고 그 값을 이용해 F-measure를 계산하였다. 논쟁 쌍 검출 결과는 [그림 12]와 같다. 논쟁댓글 검출에서와 마찬가지로 두 사이트 모두 유효논쟁지수의 임계치가 높을수록 정확도는 증가하고, 재현율은 감소하는 것으로 나타난다.

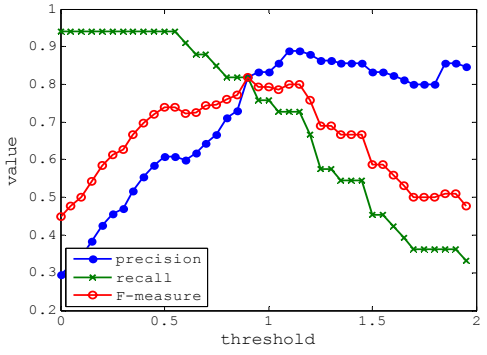
[표 9]는 각 측정치에 대한 정확한 값을 보여준다.

표 9. 제안 방법의 논쟁 작성자 쌍 검출 성능

site	θ	precision	recall	F-measure
SkepticalLeft	0.80	0.76	0.74	0.75
아고라	0.90	0.82	0.82	0.82



(a) SkepticalLeft



(b) 아고라

그림 12. 유효논쟁지수 임계치 변화에 따른 논쟁 작성자 쌍 검출 성능

SkepticalLeft에서는 임계치 0.8일 때 정확도가 0.76, 재현율이 0.74, 그리고 F-measure가 0.75로 나타났고, 아고라의 경우는 임계치가 0.9일 때 세 값이 모두 0.82로 측정되었다. 논쟁댓글 검출에서와는 달리 SkepticalLeft 사이트에서 논쟁 쌍 검출 성능이 낮게 나타났다. 그 원인으로는 SkepticalLeft의 경우 아고라보다 논쟁에 참여한 작성자들이 많아 여러 그룹의 논쟁이 겹쳐서 나타나는 경우가 많다. 이런 경우에는 실제로 서로 다른 작성자들과 논쟁하는 경우라도 인접하게 나타날 확률이 높아 논쟁 쌍으로 검출될 가능성이 높으므로 검출 성능이 낮아질 수 있다.

마지막으로 제안 방법을 이용하여 대규모 댓글 집합에서의 논쟁구간 및 논쟁관계를 시각화하였다. 실험 데이터 집합에서 댓글이 많은 두 개의 게시물을 선택하고 각 게시물에 포함된 댓글 집합을 대상으로 논쟁구간과

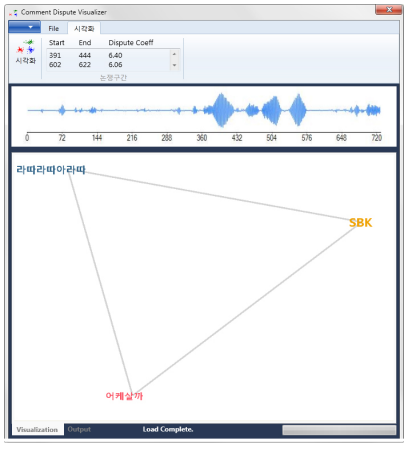
논쟁 쌍에 대한 시각화를 수행하였다. [표 10]은 실험 게시물의 정보를 정리한 것이다.

표 10. 논쟁관계 검출 실험을 위한 게시물

게시물	등록일	댓글수	작성자수
#3188361	09.12.01	847	493
#3219398	09.12.18	826	465

두 개의 게시물 모두 아고라 게시판에 등록된 것으로 댓글수가 모두 800개 이상으로 게시판 이용자들의 관심을 많이 받은 게시물들이다. 각 게시물에 대한 논쟁구간 및 논쟁관계 시각화 결과는 [그림 13]에서 보여준다.

(a) 게시물 #3188361



(b) 게시물 #3219398

그림 13. 아고라 게시물의 논쟁구간 및 논쟁관계 시각화 결과

[그림 13]에서 윈도우 상단에 검출된 논쟁구간이 논쟁지수와 함께 나타나고 전체 댓글의 논쟁 지수가 파형 그래프로 나타난다. 이때 진폭이 제일 큰 구간이 최대 논쟁구간이다. 윈도우 하단의 비방향성 그래프는 논쟁구간 내에서 논쟁관계에 있는 작성자들의 관계를 보여준다. 이 논쟁관계 그래프에서 노드는 작성자 ID를 가리킨다. [그림 13](a)의 논쟁관계를 통해 ID가 '격랑의 바다'인 작성자가 나머지 5명의 작성자와 논쟁하고 있는 것으로 판단할 수 있다. 이 경우처럼 한 작성자가 많은 작성자와 논쟁관계를 가지는 경우로 나타날 때는 정상적인 논쟁으로 보기 어렵다. 실제 댓글의 내용을 살펴본 결과 해당 작성자의 댓글은 논쟁과는 관계없이 다른 사람의 댓글에 계속해서 억지를 쓰는 댓글임을 알 수 있었다. 한 게시물에 같은 댓글을 계속해서 다는 스팸 댓글도 이런 논쟁관계로 시각화될 수도 있을 것이다.

VI. 결론

본 논문에서는 인터넷 토론 게시판에서 논쟁댓글이 지니는 특성을 분석하였다. 우리는 일반적인 오프라인에서 벌어지는 토론이나 논쟁과 같이 논쟁댓글이 지니는 다음과 같은 특징에 착안하였다. 첫 번째는 논쟁중인 작성자들은 그렇지 않은 작성자들보다 한 게시물에서 댓글을 작성하는 빈도가 높다. 두 번째는 논쟁중인 작성자의 댓글은 자신이 이전에 쓴 댓글과 인접한 거리에 다시 나타날 확률이 높으며, 논쟁중인 작성자들의 글이 서로 번갈아가며 나타날 확률이 높다. 이러한 가정을 검증하기 위해서 실제 인터넷 토론 게시판인 SkepticalLeft와 아고라 사이트의 게시물의 댓글을 수집하고 이 데이터를 분석하였다. 분석 결과 댓글 및 댓글 작성자들의 분포에서 논쟁중인 작성자들의 댓글 작성 빈도가 높고, 논쟁에 참여한 작성자들의 댓글거리가 짧게 나타나 우리의 가정이 올바름을 알 수 있었다.

또한 본 논문에서는 이러한 논쟁댓글의 특성을 바탕으로 댓글목록에서 논쟁이 벌어지는 구간을 검출하고, 논쟁구간 내의 논쟁 작성자 쌍을 검출하였다. 논쟁구간을 식별하기 위해 각 작성자들의 댓글 거리를 이용하여

댓글의 논쟁지수를 구하고 논쟁지수의 합이 가장 큰 구간을 논쟁구간으로 검출하였다. 그리고 논쟁구간 내에서 어떤 작성자들이 논쟁관계가 있는지를 알아보기 위해서 작성자 쌍의 논쟁지수를 구하고 임계치 이상의 논쟁지수를 가지는 작성자 쌍을 논쟁 쌍으로 검출하였다. 그리고 검출된 논쟁구간 및 논쟁관계를 꺾은 선 그래프와 비방향성 그래프로 시각화하여 댓글에 포함된 논쟁구조를 쉽게 파악할 수 있도록 하였다.

제안 방법의 성능을 보이기 위해 인터넷 토론 사이트인 SketpicalLeft와 아고라 사이트에서 수집한 댓글을 이용해서 논쟁댓글과 논쟁 쌍을 검출하고 정확도와 재현율 그리고 F-measure를 측정하였다. 논쟁댓글 검출 성능은 F-measure가 0.84 (SkepticalLeft)와 0.83 (아고라)으로 측정되었고, 논쟁 쌍 검출은 각각 0.75 (SketpcialLeft)와 0.82 (아고라)로 측정되었다.

제안 방법은 댓글 작성자들의 순서관계를 이용하여 댓글 목록에 포함된 논쟁을 검출하므로 사용 언어에 제약 없이 적용할 수 있다. 또한 하나의 뷰로 시각화된 결과를 통해 쉽고 직관적으로 논쟁관계를 파악할 수 있도록 해줌으로써 인터넷 토론 게시판 이용자에게 편리한 도구로 사용될 수 있을 것이다.

참 고 문 헌

[1] C. Marlow, “Audience, structure and authority in the weblog community,” In The 54th Annual Conference of the International Communication Association, 2004.

[2] 심재민, 조찬형, 양효진, 안인회, 나은아, *웹 2.0 시대의 네티즌 인터넷 이용 현황*, 한국인터넷진흥원, 2006.

[3] http://forums.canadiancontent.net/

[4] http://politics.conforums.com/

[5] G. Mishne and N. Glance, “Leave a reply: An analysis of weblog comments,” In Third annual workshop on the Weblogging ecosystem. Citeseer, 2006.

[6] Y. Lee., M. Bae, G. Woo, and H. Cho, “A Personalized Visualizing and Filtering system for a Large Set of Responding Messages on Internet Discussion Forums,” In Proc. of the CIT09, Vol.2, pp.160-165, 2009.

[7] 이윤정, 지정훈, 우균, 조환규, “인터넷 게시물의 댓글 분석 및 시각화”, 한국콘텐츠학회논문지, 제9권, 제7호, pp.45-56, 2009.

[8] 이윤정, 김은경, 조환규, 우균, “스킵리스트를 이용한 인터넷 토론 게시판 댓글 관리”, 한국콘텐츠학회논문지, 제10권, 제8호, pp.38-50, 2010.

[9] M. Hu, A. Sun, and E. Lim, “Comments-oriented blog summarization by sentence extraction,” Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp.901-904, 2007.

[10] L. Zhou and E. Hovy, “On the summarization of dynamically introduced information: Online discussions and blogs,” Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs, Standford, 2006.

[11] T. Kamada and S. Kawai, “An algorithm for drawing general undirected graphs,” Information processing letters, Vol.31, No.1, pp.7-15, 1989.

[12] http://skepticalleft.com

[13] http://agora.media.daum.net

저 자 소 개

이 윤 정(Yun-Jung Lee) 정회원

- 1995년 2월 : 부경대학교 전자계산학과(이학사)
- 1999년 2월 : 부경대학교 전산정보학과(이학석사)
- 2008년 8월 : 부경대학교 전자계산학과(이학박사)

▪ 2008년 9월 ~ 현재 : 부산대학교 U-Port 정보기술사업단 박사후연구원
<관심분야> : 얼굴 애니메이션, 웹 콘텐츠 시각화

정 인 준(In-Joon Jung) 준회원

▪ 2010년 8월 : 부경대학교 컴퓨터멀티미디어공학과(이학사)
▪ 2010년 9월 ~ 현재 : 부산대학교 컴퓨터공학과 석사과정
<관심분야> : 데이터 클러스터링, 코드 시각화

우 균(Gyun Woo) 정회원

▪ 1991년 : 한국과학기술원 전산학(학사)
▪ 1993년 : 한국과학기술원 전산학(석사)
▪ 2000년 : 한국과학기술원 전산학(박사)
▪ 2000년 ~ 2002년 : 동아대학교 컴퓨터공학과 전임강사
▪ 2002년 ~ 2004년 : 동아대학교 컴퓨터공학과 조교수
▪ 2004년 ~ 현재 : 부산대학교 컴퓨터공학과 부교수
<관심분야> : 프로그래밍언어 및 컴파일러, 함수형 언어, 그리드컴퓨팅, 소프트웨어 메트릭, 프로그램 시각화