인터넷 게시물의 댓글 분석 및 시각화

Analysis and Visualization for Comment Messages of Internet Posts

이윤정, 지정훈, 우균, 조환규
부산대학교 컴퓨터공학과

Yun-Jung Lee(leeyj01@pusan.ac.kr), Jeong-Hoon Ji(jhji@pusan.ac.kr),
Gyun Woo(woogyun@pusan.ac.kr), Hwan-Gue Cho(hgcho@pusan.ac.kr)

요약

오늘날 인터넷 사용자들은 블로그나 뉴스, 인터넷 게시판 등의 매체에서 댓글을 통해 다른 사람의 의견을 살피고 자신의 의견을 나타내고 있다. 그러나 현재 대부분의 블로그나 인터넷 포털 사이트의 경우 기사나 댓글들을 순차적인 목록 형태로 제공하므로 사용자가 원하는 내용의 댓글을 검색하거나 살펴보는 것은 힘든 일이다. 또한 댓글 사용자가 증가함에 따라 스팸 댓글이나 악플 등이 사회 문제가 되기도 한다. 본 논문에서는 다음 아고라(Daum AGORA) 웹 블로그의 게시글과 댓글을 통계적으로 분석하고 유사도를 기반으로 클러스터링하는 시스템을 제안한다. 본 시스템은 클러스터링 결과를 시각화하여 간단한 스크린 뷰(screen view)로 보여준다. 또한, 본 시스템은 생물정보학에서 잘 알려진 정렬 기법인 Needleman-Wunsch 알고리즘을 이용해 스팸 댓글을 필터링한다.

■ 중심어 : | 블로그 시각화 | 인터넷 게시물 | 댓글 | 스팸 필터링 |

Abstract

There are many internet users who collect the public opinions and express their opinions for internet news or blog articles through the replying comment on online community. But, it is hard to search and explore useful messages on web blogs since most of web blog systems show articles and their comments to the form of sequential list. Also, spam and malicious comments have become social problems as the internet users increase. In this paper, we propose a clustering and visualizing system for responding comments on large-scale weblogs, namely 'Daum AGORA,' using similarity analysis. Our system shows the comment clustering result as a simple screen view. Our system also detects spam comments using Needleman-Wunsch algorithm that is a well-known algorithm in bioinformatics.

■ keyword : | Blog Visualization | Internet Posts | Comment Messages | Spam Filtering |

I. 서 론

사이버 공간(cyber space)은 다른 사람들과 공유하면서 집단적으로 구성되는 네트워크상의 가상세계로, 일종의 물리적 공간인 동시에 사회적 공간이라고 할 수 있다. 특히 웹 2.0을 기반으로 하여 블로그, 인터넷 포럼 등과 같은 쌍방향 의사소통이 가능한 다양한 형태의 온라인 커뮤니티를 형성하여 새로운 여론 형성의 장이 되

접수번호 : #090706-002
접수일자 : 2009년 07월 06일
심사완료일 : 2009년 07월 20일
교신저자 : 우균, e-mail : woogyun@pusan.ac.kr

고 있다. 인터넷을 통한 의사소통은 게시판, 온라인 카페, 커뮤니티, 포털 등 다양한 형태로 이루어지며 게시글 쓰기와 읽기뿐만 아니라 댓글 달기, 100자 의견, 채팅, 여론조사 참여, 글 퍼 나르기, 사진 올리기, 음악 만들기 등과 같은 행위를 포괄적으로 담고 있다. 인터넷 이용자가 손쉽게 접근할 수 있는 대부분의 포털 사이트의 경우 자유게시판, 커뮤니티 게시판, 뉴스 게시판 등 다양한 형태의 인터넷 게시판을 운영하고 있다.

댓글의 경우 해당 게시물에 대한 가장 적극적인 의사표현 형태로 볼 수 있으며, 게시물과 게시물 독자들과 상호작용할 수 있는 쉽고 효율적인 방법으로 사용되고 있다[1]. 댓글이 달렸다는 사실은 다른 사람들도 그 뉴스에 주목하였다는 점을 가시적으로 알 수 있게 해 주는 일종의 신호(signal) 역할을 한다. 사람들이 댓글을 보든 보지 않든 댓글 자체가 존재함으로써 기사는 좀 더 주목할 만하고 중요한 것으로 인식될 수 있다[2]. 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[3]. 실제로 '아고라'의 경우 게시물에 대한 의견을 300자 이내의 짧은 댓글을 이용해 실시간 논쟁이 일어나는 경우를 흔히 볼 수 있다.

대부분의 인터넷 게시판이나 블로그 등에서는 게시물에 댓글을 달 수 있게 허용하고 있으나 게시물과 마찬가지로 목록 형태로 제공하고 있어 많은 수의 댓글이 달릴 경우 전체적인 댓글 내용을 파악하기 힘들고, 검색이나 정렬 기능이 제공되지 않아 댓글의 임의 접근이 어려운 실정이다. 최근 인터넷 공간에서 댓글 사용이 증가함에 따라 익명성을 이용하여 게시물의 내용과 관련 없는 광고성 스팸 댓글이나 악성댓글로 인한 피해도 증가하고 있다. 이러한 스팸이나 악성 댓글의 차단을 위해 여러 가지 방법들이 제안되고 있으나 그 효과는 아직 미흡한 실정이다[4-6].

본 논문에서는 인터넷 게시판이나 블로그의 게시물에 포함된 댓글을 내용에 따라 분류하고 이를 시각화하는 시스템을 제안한다. 제안 시스템에서는 댓글을 내용의 유사도에 따라 여러 클러스터로 분류하고, 분류 결과를 하나의 뷰로 시각화함으로써 이용자들이 게시물에 포함된 댓글에 대한 성향을 쉽게 파악할 수 있도록 해주며, 자신이 원하는 내용을 포함한 댓글을 선택적으로 접근할 수 있게 해준다. 또한 동일하거나 비슷한 내용이 반복적으로 나타나는 스팸 댓글 및 의미 있는 단어가 포함되어 있지 않은 무의미한 댓글을 추출하여 다른 댓글들과 구분해 줌으로써 이용자들에게 불필요한 댓글이 노출되지 않도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 댓글 분석과 댓글 시각화에 관한 기존 연구들을 살펴본다. 3장에서는 제안 시스템의 개요와 데이터 집합에 대해 설명한다. 4장과 5장에서는 제안 시스템의 댓글 분류 및 시각화에 대해 설명한다. 마지막으로 6장에서 실험 결과를 보이고, 7장에서 결론을 맺는다.

II. 관련 연구

인터넷 게시물의 댓글에 초점을 맞춘 연구로 Herring[7]의 연구를 들 수 있다. 203개의 블로그를 대상으로 한 조사에서 블로그 게시물 당 평균 0.3개의 댓글이 작성되는 것으로 나타났다. 그러나 전체 블로그 공간으로 확장해 명확한 결론을 내기에 데이터 샘플이 너무 적고, 댓글 자체에 대해 더 이상 분석되어 있지 않다. Trevino와 Gumbrecht의 연구에서는 블로그의 인터랙티브한 속성을 위해 댓글이 필수적임을 제시하였다[8,9]. Krishnamurthy의 연구에서는 특정 블로그에서 '911 사건'에 대한 게시물들의 패턴을 조사한 결과 통찰력 있는 게시물일수록 많은 댓글이 달리는 것으로 나타났다[10]. Mishne과 Glance[11]는 상당히 큰 댓글 집합을 사용하여 블로그 공간에서에 차지하는 댓글의 비중을 추정하고, 블로그의 인기와 댓글 패턴간의 관계를 분석하였다. 이 연구에서 블로그 공간에서 댓글은 블로그 게시물 양의 약 30% 정도로 상당한 비중을 차지하고 있으며, 댓글의 양은 블로그나 게시물에 대한 관심 정도를 가리키는 지시자로 사용될 수 있음을 제안하였

다. 또한 댓글의 내용을 분석하여 논쟁 정도를 계산하는 새로운 방법을 제시하였다.

G. Mishne[12] 등은 블로그에서 스팸 댓글 분류 방법을 제안하였다. 이 연구에서는 언어 모델을 이용하여 블로그의 본문과 댓글, 댓글이 링크된 페이지간의 유사도 비교를 통해 스팸 여부를 판단한다. 이와 유사한 연구로 배민영[4]은 문서요약이나 문서분류에 사용하는 토픽 시그너처(topic signature)를 이용하여 악성 댓글을 분류하는 시스템을 제안하였다. 이 연구에서는 문장의 길이가 비교적 짧고, 띄어쓰기가 거의 없으며, 특수기호가 많이 사용되는 등의 악성 댓글이 가지는 특징을 이용한 패턴 매칭 방법을 통해 악성 댓글의 분류 성능을 개선할 수 있음을 보였다.

Harris[13] 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. 이 시스템은 많은 기사들을 표현하고 있으나 어떤 블로그에서 게시되었는지 혹은 기사들의 앞, 뒤 연결을 알 수 없다.

이것과 유사하게 BBC에서는 뉴스에 대한 댓글을 시각화하는 시스템인 Spectrum을 개발하였다[14]. BBC 2's White 시즌 중 토론을 조사하여 감정, 지역, 성별 등에 따라 댓글을 클러스터링하고 이를 시각화한다. 각 댓글들은 감정 별로 분류되어 서로 다른 색의 원으로 시각화 된다. Spectrum은 감정, 지역, 성별 등과 같이 분류할 기준을 선택할 수 있는 사용자 인터페이스를 제공하고 있어 원하는 기준으로 댓글들을 필터링할 수 있고 움직이는 입자를 클릭하면 토론에서 사용된 댓글을 볼 수 있다. 그러나 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[15] 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. iBlogVis에서는 블로그 내의 게시물과 댓글의 전체적인 현황은 파악할 수 있으나 리스트로 제공되는 것과 마찬가지로 댓글의 검색이나 게시물과의 의미적 관계 등은 파악할 수 없다.

블로그 게시물이나 블로그 자체에 대한 연구에 비해 댓글의 분석 및 시각화에 대한 연구는 상대적으로 그 수가 적으며, 댓글 자체에 대한 분석 및 연구가 부족한 실정이다. 비교적 최근 연구로서 댓글 시각화 시스템 TRIB[16]가 발표된 바 있다. 본 논문은 이 연구의 후속 연구로서 스팸 및 무의미 댓글을 필터링 할 수 있는 댓글 분석 및 시각화 시스템을 제안한다. 한편 클러스터링을 이용하여 스팸 필터링을 수행하는 연구도 이미 진행된 바 있다[17][18]. 본 연구에서는 생물정보학의 정렬 기법[19]을 이용하여 스팸 필터링을 수행한 후에 클러스터링 기법을 이용하여 댓글을 분석하고 시각화하는 기법에 대하여 소개한다.

III. 제안 시스템 및 데이터 수집

1. 제안 시스템의 개요

본 논문에서는 게시물에 달린 댓글들을 내용에 따라 분류하여 댓글 전체에 대한 개관을 제시하고, 댓글에 포함된 스팸이나 논쟁과 같은 특성을 찾아내어 이를 시각화하는 시스템을 제안한다. 제안 시스템의 구성은 [그림 1]과 같다.

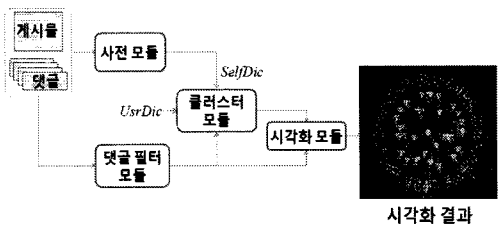


그림 1. 시스템 구조

제안 시스템은 사전 모듈, 댓글 필터 모듈, 클러스터 모듈, 그리고 시각화 모듈로 구성된다. 사전 모듈에서는 댓글에 포함된 명사들을 추출하여 단어 목록을 만들고, 목록에 포함된 단어 중에서 발생 빈도가 높은 상위 25개의 단어를 이용하여 댓글 사전인 *SelfDic*을 생성한다. 또한 댓글의 내용과는 관계없이 이용자가 관심을 가지는 단어들을 모아 사용자 정의 사전인 *UsrDic*을 생성한다. 두 사전에 정의된 단어들은 댓글 분류 시 키워드로 사용되며 각 단어별로 댓글 클러스터를 형성한다. 클러스터 모듈에서는 댓글의 내용과 키워드 간의 유사도를 계산하여 가장 높은 유사도를 가지는 키워드 클러스터로 해당 댓글을 분류한다. 댓글 필터 모듈은 전체의 댓글에서 스팸이나 무의미 댓글을 필터링 한다. 스팸 댓글은 동일하거나 비슷한 내용의 댓글이 반복적으로 나타나는 것을 의미하고, 무의미 댓글은 내용에 의미 있는 단어가 포함되어 있지 않고 의성어나 비속어 등이 나타나는 댓글을 의미한다. 마지막으로 시각화 모듈에서는 키워드로 분류된 댓글 및 스팸과 무의미 댓글들을 서로 다른 색상의 원으로 나타내어 간단한 스크린뷰를 생성한다.

2. 데이터 수집

블로그와 같이 인터넷 이용자들의 자발적인 게시물 작성과 의견 교환이 이루어지는 인터넷 공간에서 얼마나 많은 댓글들이 생성되고 있는지를 살펴보기 위해 다음 포털의 아고라 게시판의 게시물들을 수집, 조사하였다. 아고라는 정치, 경제, 사회, 문화 등 다양한 분야의 게시물들이 등록되는 토론 게시판으로 이용자들이 자유롭게 게시물을 작성하거나 다른 사람들의 게시물을 읽을 수 있으며, 찬성/반대와 댓글을 이용해 게시물에 대한 자신의 의견을 나타낼 수 있다.

[표 1]은 이용자들의 관심이 높은 게시물이 등록되는 '토론베스트'에 등록된 게시물의 현황을 보여준다. 2009년 1월 1일에서 2009년 4월 30일까지 '토론베스트'에 등록된 게시물을 조사한 것으로, 4개월 동안 총 게시물 수는 54,410개로 하루 평균 약 453개의 게시물이 등록되었다. 그리고 게시물 당 평균 조회수는 약 2,623건, 평균 댓글수는 26개, 그리고 조회수에 대한 댓글 비는 약 1%로 나타났다. '토론베스트' 게시물의 경우 이용자들의 관심이 높은 게시물로 조회수가 수십에서 수만 건 이상으로 범위가 넓게 나타난다.

표 1. 2009년 1사분기 '다음 아고라 '의 ' 토론 베스트 ' 게시물 현황

날짜	게시물	조회 수	댓글 수	비율
09.01.01~09.01.31	20,126	51,586,079	512,356	0.99%
09.02.01~09.02.28	16,116	39,177,161	390,993	1.00%
09.03.01~09.03.31	10,837	31,899,492	267,640	0.84%
09.04.01~09.04.30	7,331	20,073,066	224,892	1.12%
합계	54,410	91,149,719	883,525	0.97%
게시물 당 평균		2,623	26	0.99%

[그림 2]는 조회수 구간별 게시물 분포를 보여준다. 조회수 500건 이하의 게시물이 전체 게시물 54,410개 중 25,015개인 48%로 가장 많고, 5,000건 이상의 조회수를 가지는 게시물도 7,934개로 전체 15%를 차지한다.

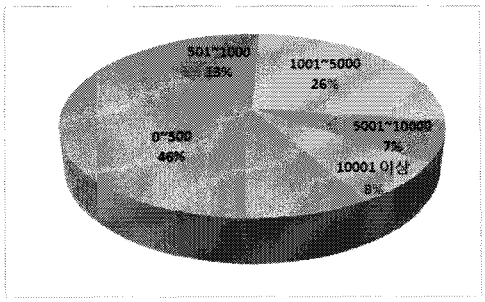


그림 2. '다음 아고라' 의 조회수 구간별 게시물 분포

많은 조회수를 가지는 게시물들은 지속적으로 이용자의 관심을 끌게 되고, 많은 댓글이 달리는 등 인터넷 게시판 상의 여론을 만드는 역할을 한다고 볼 수 있다.

[표 2]는 조회수 구간별 게시물 수와 댓글수 및 평균 댓글수를 보여준다.

표 2. 조회수 구간별 게시물 수와 댓글 수

조회수 구간	게시물	댓글 수	평균 댓글 수
0~500	25,015	135,641	5.42
501~1,000	7,145	93,335	13.06
1,001~5,000	14,316	339,005	23.68
5001~10,000	3,888	198,809	51.13
10,001 이상	4,046	629,091	155.48
전 체	54,410	1,395,881	25.65

조회 수가 10,000건 이하인 대다수의 게시물들은 평균 댓글 개수가 약 50개 이하로 나타났으나, 조회 수가 10,000건 이상인 약 4,046 개의 게시물들은 총 댓글 수는 375,279개, 게시물 당 평균 댓글 수는 약 155개로 조사되었다.

[표 3]은 댓글 개수에 따른 게시물 수를 보여준다. 전체 게시물의 약 90% 정도가 50개 이하의 댓글을 가지고 있으며, 100개 이상의 많은 댓글을 가지는 게시물은 2,144개인 약 4% 정도로 나타났다.

표 3. 댓글수 구간별 게시물 현황

댓글수	게시물수	비율
0~50	49,082	90.2%
51~100	3,184	5.8%
101~500	1,886	3.5%
501~1,000	167	0.3%
1,001 이상	91	0.2%
합계	54,410	100.0%

많은 수의 댓글을 가지는 게시물의 비중은 적지만 많은 이용자에게 핫이슈가 되고 있는 게시물로 간주될 수 있으며, 이러한 게시물의 경우 전체 댓글을 읽거나 다수의 의견을 수렴하는데 어려움이 있다.

IV. 댓글 분류

1. 사전 정의

제안 시스템에서 댓글은 두 가지 방법으로 분류될 수 있다. 그 중 하나는 게시물의 댓글에서 많이 나타나는 단어들을 기준으로 분류하는 것이고, 또 다른 방법은 이용자가 관심 있는 단어들을 이용하는 방법이다. 두 방법 모두 선택된 단어들을 이용하여 사전을 구성하고, 각 단어들을 키워드로 하여 댓글 내용과 가장 유사도가 높은 단어 클러스터로 댓글을 분류하게 된다. 전자의 경우를 댓글 사전이라고 하고, 후자의 경우는 사용자 정의 사전이라고 한다. 댓글 사전을 이용하여 댓글을 분류 할 경우 많은 댓글에 대한 전체적인 내용을 직관적으로 파악할 수 있으며, 사용자 정의 사전을 이용하는 경우에는 이용자가 관심을 가지는 내용을 포함하는 댓글들만 선택할 수 있다는 장점이 있다.

댓글 사전을 구성하기 위해 각 댓글마다 띄어쓰기를 단위로 분리하여 단어 목록을 만든다. 이 단어 목록에서 미리 정의된 표준 명사 사전에 있는 단어들만 추출한 후, 가장 많은 댓글에서 나타나는 상위 25개의 단어들만 선택하여 댓글 사전을 구성한다.

2. 댓글 분류

댓글 분류를 위해 먼저 사전에 정의된 키워드와의 댓글 내용과의 의미적 유사도를 계산한다. 키워드 t_i와 댓글 c_k의 의미적 유사도인 $w(t_i, c_k)$는 식 1을 통해 구할 수 있다.

$$w(t_i, c_k) = f_{i,k} / \sum_{j=1}^{|T|} f_{j,k}, \quad \forall_{t_i} \in T \qquad (1)$$

여기서 T는 사전을 의미하고, $f_{i,k}$는 t_i가 댓글 c_k에서 나타나는 빈도수이다.

댓글은 전체 키워드와의 의미적 유사도를 계산하여 가장 큰 의미적 유사도를 가지는 키워드 클러스터로 분류된다. 만일 동일한 유사도를 가지는 키워드들이 있을 경우 댓글은 그 키워드들의 클러스터 중에서 적은 수의 댓글을 가지는 클러스터에 속하게 된다. 이것은 한 클러스터에 너무 많은 댓글이 집중되는 것을 방지하기 위한 것이다.

어떤 키워드도 포함하지 않는 댓글의 경우는 키워드 클러스터로 분류되지 않고, 시각화 과정에서 댓글 작성자 id에 따라 배치된다.

3. 스팸 및 무의미 댓글

많은 수의 댓글 중에는 동일한 작성자가 같은 내용을 반복적으로 올리는 스팸 댓글이나 의미 있는 단어를 포함하지 않는 무의미 댓글들을 어렵지 않게 찾아볼 수 있다. 현재의 순차적 댓글 접근 방법에서는 이렇게 불필요한 댓글이 이용자가 원치 않아도 다른 댓글들을 읽다보면 이용자에게 노출될 수밖에 없다. 제안 시스템에서는 댓글 내용을 분석하여 스팸이나 무의미한 댓글들을 따로 분류하여 시각화함으로써 이용자들에게 불필요한 댓글이 노출되는 것을 피하도록 한다.

먼저 스팸 댓글 분류를 위해서 같은 작성자가 작성한 댓글간의 문자열 유사도를 구하고, 임계치 이상의 유사도를 가지는 댓글들은 스팸 댓글로 간주한다. 제안 시스템에서는 댓글의 유사도 분석을 위해 두 문장에 대한 전역 정렬(global alignment)를 수행하는 Needleman-Wunsch 알고리즘을 이용한다[19]. [그림 3]과 [그림 4]는 스팸 댓글의 예를 보여준다.

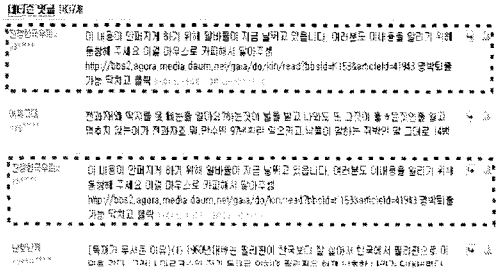


그림 3. 동일한 내용을 반복적으로 작성한 스팸 댓글

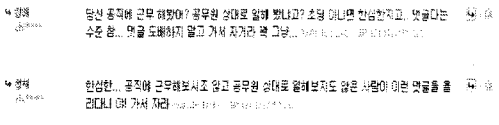


그림 4. 비슷한 내용을 반복적으로 작성한 스팸 댓글

[그림 3]은 동일한 내용을 반복적으로 작성한 스팸 댓글을 나타내고, [그림 4]는 동일하지는 않으나 거의 비슷한 내용을 반복적으로 작성한 스팸 댓글을 나타낸다.

다음으로 의미 있는 단어를 포함하지 않는 무의미한 댓글을 추출하기 위해 댓글 사전 생성에서와 같이 댓글 내용을 단어로 분리하고 표준 명사 사전에 포함되는 단어들만 추출한다. 이 때 추출된 단어가 하나도 없는 댓글을 무의미 댓글로 간주한다. 무의미 댓글의 예가 [그림 5]에 나타나 있다.

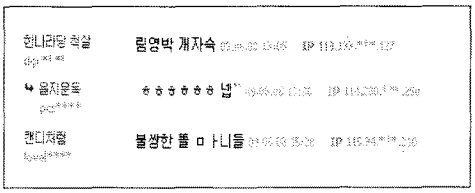


그림 5. 무의미 댓글의 예

제안 시스템에서는 욕설이나 기타 부적절한 표현을 포함하는 악성 댓글보다 좀 더 포괄적인 개념으로 무의미 댓글을 분류함으로써 악성 댓글뿐만 아니라 의성어, 기호 등으로 구성된 댓글처럼 이용자가 읽을 필요가 없는 댓글까지 함께 분류한다.

2009년 1월 10일 '아고라'의 '자유토론'에 등록된 게시물 중 조회수가 120,302건인 게시물의 경우 1,837개의 댓글 중에서 스팸 댓글이 120개, 무의미 댓글이 308개로 분류되었다. 무의미 댓글 중 38개의 댓글은 스팸 댓글이다.

V. 시각화

제안 시스템에서 시각화된 뷰는 태양계와 유사한 구조로 배치되어 있다. 화면 중심에 게시물을 두고 그 주변으로 사전에 속한 키워드가 배치되며, 각 키워드의 클러스터로 분류된 댓글들은 키워드를 중심으로 방사형으로 배치된다. 다시 말해서, 게시물이 태양에 해당하고 키워드는 행성에 해당하며 댓글들은 행성에 속한 위성처럼 배치된다. 키워드 및 댓글 배치방법은 TRIB와 같다[16].

[그림 6]은 제안 시스템의 키워드와 댓글 배치를 보여준다.

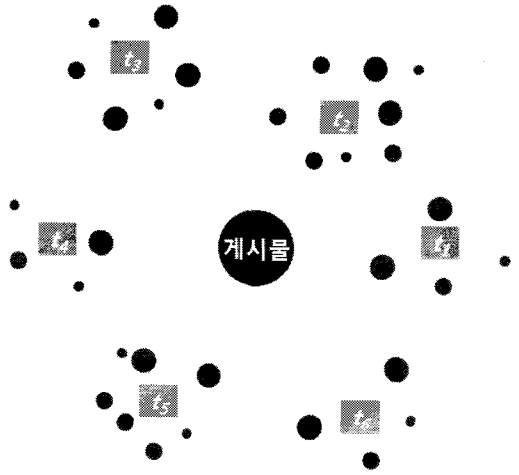


그림 6. 키워드 및 댓글의 배치

그림은 6개의 키워드와 키워드로 분류된 댓글을 가상으로 배치한 화면이다. 제안 시스템에서는 색상이 다른 원을 이용하여 각각의 댓글의 특성을 나타낸다. 색상과 모양이 다른 시각화 요소들이 [표 4]에 나타나 있다.

표 4. 시각화 표현 요소

시각화 요소	표현 대상
	키워드
	단어로 분류된 댓글
	단어로 분류되지 않은 댓글
	스팸 댓글
	무의미 댓글
	마우스로 선택한 댓글
	선택 댓글의 이전 댓글
	선택 댓글의 이후 댓글
	10개 이상의 댓글 작성자 id

회색의 사각형은 사전에 속한 키워드를 나타내고, 각각 다른 색상을 가지는 원들은 댓글을 나타낸다. 댓글은 색상에 따라 녹색은 키워드로 분류된 댓글을 나타내고, 흰색은 키워드로 분류되지 않은 댓글, 붉은색은 스팸 댓글, 회색은 무의미 댓글을 나타낸다. 그리고 마우스로 선택한 댓글은 보라색 이중 원으로 표현하고, 댓글이 작성된 시간을 기준으로 선택된 댓글의 이전 댓글은 파란색 이중 원, 이후 댓글은 녹색 이중 원으로 시각화하여 순차적 댓글 접근도 가능하게 한다. 댓글을 표현한 원형의 크기는 댓글 글자 수에 비례한다. 마지막으로 노란색의 사각형은 10개 이상의 댓글을 작성한 id를 나타내어 반복적으로 많은 댓글을 작성하는 작성자 id를 쉽게 검색할 수 있도록 한다.

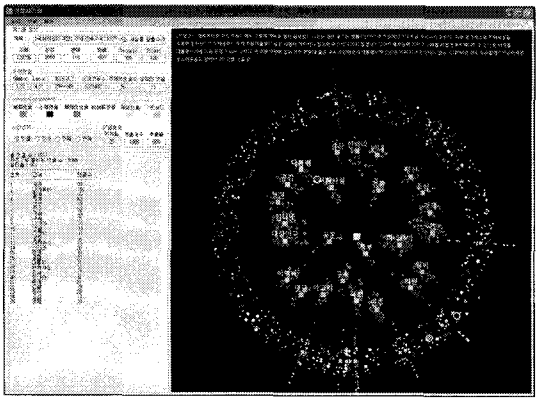


그림 7. 제안 시스템의 시각화

[그림 7]은 아고라에서 수집한 게시물 중 1,837 개의 댓글을 가지는 게시물을 제안 시스템으로 시각화한 결과이다. 그림에서 왼쪽 패널에 작성자, 조회수, 추천수, 댓글수 등과 같은 게시물에 대한 정보와 댓글 분류에 사용된 사전의 종류 및 사전에 포함된 키워드가 나타난다. 오른쪽 시각화 뷰에는 키워드로 분류된 댓글들이 앞서 설명된 시각화 요소로 표현되고, 각각의 요소를 마우스로 클릭하면 화면 상단에서 댓글 내용을 확인할 수 있으며, 화면 중앙의 파란색 지구 모양의 아이콘을 클릭할 경우 게시물이 등록된 사이트에 접속하여 해당 게시물을 확인할 수 있다. 그리고 시각화 뷰 가장자리에 원형으로 배치된 댓글들은 키워드 클러스터에 포함되지 않는 댓글로서 작성자를 기준을 배치된다. 동일한 작성자가 많은 댓글을 올린 경우에는 댓글들이 일직선에 배치됨으로써 시각화 화면에서 쉽게 식별할 수 있다.

Ⅵ. 실험 및 결과

제안 시스템은 C#으로 구현되었으며, 시각화 모듈을 위해 오픈 소스 프로그래밍 언어인 Processing이 사용되었다.

1. 댓글 분류 성능

*SelfDic*의 댓글 분류 성능을 보이기 위한 실험을 수행하였다. 실험에 사용된 게시물과 댓글은 다음 아고라 게시판의 '토론베스트'에 등록된 게시물을 수집하였다. 수집된 게시물 중에서 댓글 수가 100개 이상인 100개의 게시물을 대상으로 댓글 내용을 이용하여 *SelfDic*을 만들고 댓글을 분류하였다. 이 실험에서 *SelfDic*은 댓글 내용에 나타나는 빈도수를 기준으로 25개의 키워드로 구성되었다. 실험 데이터의 *SelfDic*의 댓글 분류 결과가 [표 5]에 나타나 있다.

표 5. *SelfDic*을 이용한 댓글 분류 비율

	댓글수	무의미 댓글	TFR	VFR
평균	221	46	65%	84%

TFR : 분류댓글수 / 댓글수 × 100%
VFR : 분류댓글수 / (댓글수-무의미댓글) × 100%

실험 데이터의 평균 댓글수는 221개, 무의미 댓글수의 평균은 46개로 조사되었다. 전체 댓글수에 대한 필터율(TFR)은 평균 65%로 나타났고, 무의미 댓글을 제외한 유효 댓글만을 고려한 경우(VFR) 평균 84%의 필터율을 보였다.

[표 6]은 댓글 분류 비율별 게시물 수를 보여준다.

표 6. 분류율 구간별 게시물 수

분류율	게시물수
0% ~ 20%	0
20% ~ 40%	0
40% ~ 60%	0
60% ~ 80%	31
80% ~ 100%	69
게시물 합계	100

80% 이상의 분류율을 가지는 게시물이 69개로 나타났으며, 60%이하의 분류율을 가지는 게시물은 보이지 않았다. 이 실험으로 제안 시스템이 많은 수의 댓글을 효율적으로 분류할 수 있음을 알 수 있다.

2. 시각화 성능

제안 시스템의 시각화 성능을 보이기 위해 1,000개 이상의 댓글을 가지는 서로 다른 주제 게시물에 대해 댓글 시각화를 수행하였다. 실험에 사용된 게시물은 [표 7]과 같다.

표 7. 댓글 시각화 실험 게시물

게시물	주제	댓글수	스팸	무의미	유효댓글
S_P	정치	1,837	120	308	1,529
S_S	사회	1,325	48	313	1,012
S_E	연예	1,336	97	438	1,400
S_G	일반	1,838	75	266	1,070

S_P와 S_S, S_E는 각각 정치와 사회, 연예에 관련된 게시물이며, S_G는 특정 주제가 없는 일반적인 게시물이다. 각 게시물의 댓글에서 스팸 댓글은 약 5.8%, 무의미 댓글은 약 20% 정도로 나타났다.

실험 게시물에 대한 제안 시스템의 시각화 실험은 *SelfDic*과 *UsrDic*을 각각 적용하여 수행되었다. *UsrDic*에 포함된 키워드는 [표 8]과 같다.

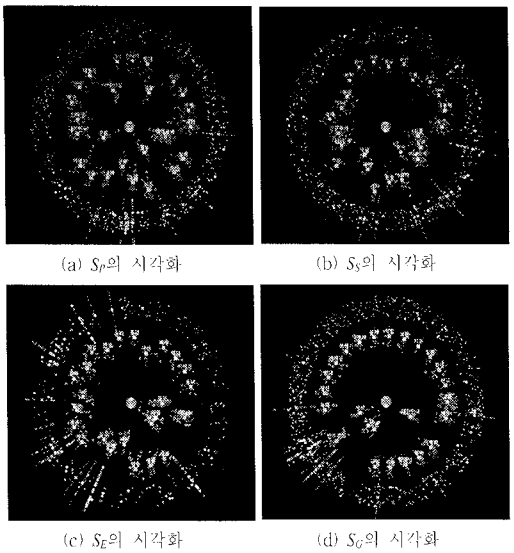
표 8. 정치 관련 *UsrDic*에 포함된 키워드

검찰	경찰	공정	구속	국가
국민	국회	노무현	대립	대통령
도덕성	미국	민주	보수	북한
언론	이명박	정권	주장	증거
집권	차별	촛불	한나라	허위

사용자 정의 사전인 *UsrDic*은 정치에 관련된 인터넷 게시물 중에서 자주 나타나는 단어들 중 25개를 임의로 선택하여 구성하였다.

[그림 9]는 4가지 게시물에 *SelfDic*을 이용하여 댓글을 분류하고 이를 시각화한 결과 화면을 보여준다. [그림 8](a)와 (b)는 각각 S_P와 S_S의 *SelfDic*을 이용하여 시

각화한 결과이고, (c)와 (d)는 S_E와 S_G의 *SelfDic*을 이용한 시각화 결과이다.



(a) S_P의 시각화 (b) S_S의 시각화

(c) S_E의 시각화 (d) S_G의 시각화

그림 8. *SelfDic*을 이용한 댓글 시각화

4가지 결과 모두 키워드 주변으로 많은 댓글이 분류되어 있는 것을 볼 수 있으며, 각 결과의 키워드를 보면 댓글에 어떤 내용이 주로 나타나는지를 직관적으로 판단할 수 있다. 또한 시각화된 결과에서 (a)의 경우 다른 게시물의 결과에 비해 키워드들이 화면 중심에 비교적 많이 몰려 있는 것을 볼 수 있다. 이것은 해당 키워드가 게시물의 본문에 많이 나타남을 의미한다.

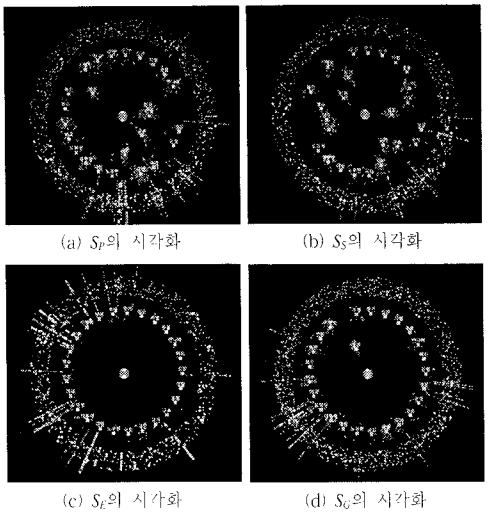
*SelfDic*을 이용한 실험 게시물의 댓글 분류 비율은 [표 9]와 같다

표 9. *UsrDic*을 이용한 댓글 분류 성능

게시물	댓글수	유효댓글	분류댓글	TFR	VFR
S_P	1,837	1,529	1,040	56.6%	68.0%
S_S	1,325	1,012	632	47.7%	62.5%
S_E	1,838	1,400	782	58.5%	73.1%
S_G	1,336	1,070	837	45.5%	59.8%

S_G의 경우가 유효 댓글만을 고려한 댓글 분류율(VFR)이 73.1% 가장 높게 나타났으며, 나머지 3개의 게시물도 약 60% 이상의 분류율을 보였다.

[그림 9]는 위의 4가지 실험 게시물에 정치에 관련된 키워드로 구성된 *UsrDic*을 적용하여 댓글을 분류하고 시각화한 결과를 보여준다.



(a) S_P의 시각화 (b) S_S의 시각화

(c) S_E의 시각화 (d) S_G의 시각화

그림 9. *UsrDic*을 사용한 댓글 시각화

그림에서 *SelfDic*을 적용했을 때에 비해 키워드 주변에 나타나는 댓글이 현저히 적음을 알 수 있다. 그중 *UsrDic*의 성격과 비교적 비슷한 S_P와 S_S의 경우 그렇지 못한 S_E와 S_G에 비해 키워드로 분류된 댓글이 많음을 볼 수 있다. 따라서 제안 시스템이 게시물의 댓글에 이용자가 원하는 내용의 댓글이 얼마나 포함되어 있음을 직관적으로 판단하고, 선택적으로 접근할 수 있도록 해준다.

[표 10]은 정치 관련 *UsrDic*을 이용한 댓글 분류 성능을 보여준다.

표 10. *UsrDic*을 이용한 댓글 분류 성능

게시물	댓글수	유효댓글	분류댓글	TFR	VFR
S_P	1,837	1,529	656	35.71%	42.9%
S_S	1,325	1,012	377	28.45%	37.3%
S_E	1,336	1,070	50	3.74%	4.7%
S_G	1,838	1,400	342	18.61%	24.4%

4개의 게시물 가운데 정치 관련 게시물인 S_P의 경우 42.9%의 분류율을 보여 가장 높게 나타났으며, S_E와 S_G

는 각각 4.7%와 24.4%와로 *SelfDic*을 이용했을 때보다 단어로 분류된 댓글의 수가 현저히 적었다.

제안 시스템을 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. [표 11]은 댓글 수를 기준으로 상위 10%의 작성자가 작성한 댓글 수에 대한 통계를 보여준다.

표 11. 댓글 수 기준 상위 10% id의 댓글 통계

기사명	평균 댓글 수	표준 편차
S_P	8.03	8.26
S_S	6.65	8.03
S_E	15.27	15.84
S_G	8.71	13.71

S_E의 경우 댓글을 많이 작성한 id의 상위 10%의 평균 작성 댓글수가 15.27개로 나머지 세 개의 기사보다 약 2배 정도로 높게 나타났다. 비교적 연예 관련 기사의 검색과 댓글 작성의 연령대가 낮은 것을 고려한다면 젊은 인터넷 이용자들의 경우 인터넷 공간에서 더 적극적으로 의사 표현을 한다고 볼 수 있다.

Ⅶ. 결론 및 향후과제

블로그와 같은 참여형 인터넷 사용이 보편화되고 이용자가 증가함에 따라 각종 블로그나 인터넷 게시판 등에는 많은 수의 게시물들과 댓글들이 게시되고 있다. 그러나 현재의 게시물들은 대부분 목록 형태로 제공되어 날짜, 작성자 등 일부 게시물에 대한 정보로 검색을 하거나 정렬을 하는 것을 제외하고는 거의 순차적 접근 밖에 허용되지 않는 실정이다. 또한 댓글의 경우는 이러한 정렬이나 검색조차도 지원되고 있지 않아 많은 수의 댓글이 달린 게시물의 경우 전체 개관을 파악하거나 자신이 원하는 내용을 선택하여 읽기가 힘들다.

본 논문에서는 인터넷 게시물에 대한 댓글 분포 및 특성을 분석하고, 내용을 기반으로 하여 댓글을 분류하고 이를 시각화하는 시스템을 제안하였다. 2009년 1월 1일에서 2009년 4월 30일까지 4달 동안 '아고라'의 '토론베스트' 게시판에 등록된 게시물을 조사한 결과 총 게시물 수는 54,410개로 하루 평균 약 453개의 게시물이 등록되었으며, 게시물 당 평균 조회수는 약 2,623건, 평균 댓글수는 26개로 조사되었다. 수집된 게시물 중 50개미만의 댓글을 가지는 게시물이 약 90%로 가장 많았으며, 500개 이상의 댓글을 가지는 게시물은 약 0.5% 정도 밖에 되지 않았으나 조회수 대 댓글비가 약 1% 정도로 조사된 것을 고려하면 이러한 게시물들은 대부분 조회수가 5만 건 이상으로 이용자들이 높은 관심을 가지는 게시물들임을 알 수 있다. 또한 id 하나 당 평균 작성 댓글 수는 4.17개로 한 작성자가 여러 개의 댓글을 작성하는 경우가 많았다. 또한 조사 게시물 중 100개의 이상의 댓글을 가지는 게시물 100개에 대하여 댓글 특성을 분석한 결과 내용을 반복적으로 게시한 스팸 댓글이 약 3%, 의미 있는 단어를 포함하지 않는 무의미 댓글은 약 23%정도로 전체 댓글에서 상당한 부분을 차지함을 알 수 있었다.

제안 시스템에서는 댓글의 분류를 위해 두 가지 종류의 단어 사전을 정의한다. 첫째는 댓글에 포함된 단어들 중에서 빈도수가 높은 단어들을 모아 *SelfDic*을 만들고, 사전에 속한 단어와 댓글 내용의 유사도를 계산하여 가장 높은 유사도를 가지는 단어 클러스터로 분류한다. 이렇게 함으로써 댓글에 주로 나타나는 내용을 파악할 수 있으며, 댓글 전체에 대한 개관을 얻을 수 있다. 두 번째로 이용자가 관심을 가지는 단어들을 이용하여 *UsrDic*을 만들고, 이것을 이용하여 댓글을 분류한다. *UsrDic*을 사용하여 댓글을 분류할 경우 많은 수의 댓글 중에서 이용자가 원하는 내용만 선택해서 읽을 수 있다. 또한 제안 시스템은 댓글 분류 결과를 시각화하여 이를 하나의 뷰로 보여준다. 댓글의 특성에 따라 서로 다른 색상의 원으로 시각화 하여 댓글의 특성 및 다수의 댓글을 작성한 id 분포를 직관적으로 파악할 수 있다.

기존의 블로그 관련 시각화 연구들은 주로 블로그 공간이나 많은 양의 게시물을 하나의 화면에 보여주기 위한 연구에 치중에 있는 반면에 본 논문에서 제안한 시

각화 방법은 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 댓글 수에 따라 가변적인 사전 구성을 생각해 볼 수 있다. 현재는 댓글 수에 관계없이 25개로 고정 키워드를 사용하고 있어 적은 수의 댓글이 달린 게시물의 경우 키워드가 하나도 나타나지 않는 댓글이 많고, 수천 개의 댓글을 가지는 경우는 25개 키워드만으로는 분류되지 않는 댓글들이 많이 생기는 것을 볼 수 있다. 따라서 댓글 수에 따라 키워드를 가변적으로 선택할 필요가 있다. 그리고 현재 사전 구성 및 의미적 유사도를 구하기 위해 댓글에서 명사만을 추출하고 있으나 더 정확한 내용 분류를 위해 문맥적 의미를 고려하는 방법도 향후 연구로 생각해 볼 수 있다.

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, pp.1-9, 2004.

[2] 김은미, 선유화, "댓글에 대한 노출이 뉴스 수용에 미치는 효과", 한국언론학보, 제50권, 제4호, pp.33-64, 2006.

[3] 심재민, 조찬형, 양효진, 안인회, 나은아, "웹2.0 시대의 네티즌 인터넷 이용 현황", 2006년 인터넷 이슈심층조사 보고서, 한국인터넷진흥원, 2006.

[4] 배민영, 차정원, "Topic Signature를 이용한 댓글 분류 시스템", 정보과학회논문지: 소프트웨어 및 응용, 제35권, 제12호, pp.774-779, 2008.

[5] http://en.wikipedia.org/wiki/Spam_in_blogs.

[6] G. Mishne and D. Carmel, "Blocking Blog Spam with Language Model Disagreement," 1st International Workshop on Adversarial Information Retrieval on the Web. pp.1-6, 2005.

[7] S. C. Herring, L. A. Scheidt, S. Bonus, and E. Wright, "Bridging the gap: A genre analysis of weblogs," In The 37th Annual Hawaii International Conference on System Sciences(HICSS'04), 2004.

[8] M. Gumbrecht, "Blogs as protected space," In WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2004 at WWW'04: the 13th international conference on World Wide Web, 2004.

[9] E. M. Trevino, "Blogger motivations: Power, pull, and positive feedback," In Internet Research 6.0, 2005.

[10] S. Krishnamurthy, "The multidimensionality of blog conversations," The virtual enactment of september 11. In Internet Research 3.0, 2002.

[11] G. Mishne and N. Glance, "Leave a reply: An analysis of weblog comments," In Third annual workshop on the weblogging ecosystem, 2006.

[12] G. Mishne, D. Carmel, and R. Lempel, "Blocking Blog Spam with Language Model Disagreement," In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web(AIRWeb), pp.1-6, 2005.

[13] http://www.wefeelfine.org.

[14] http://www.bbc.co.uk/white/spectrum.shtml.

[15] 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.

[16] 배민정, 이윤정, 지정훈, 우균, 조환규, "TRIB: 웹블로그 댓글분류 시각화 시스템", 제31회 한국정보처리학회 춘계학술발표대회 논문집, 제16권, 제1호, pp.226-229, 2009.

[17] L. Xiao-bing and N. Zhang, "Incremental

Immune-Inspired Clustering Approach to Behavior-Based Anti-Spam Technology," International Journal of Information Technology, Vol.12, No.3, pp.111-120, 2006.

[18] W.-F. Hsiao, T.-M. Chang, and G.-H. Hu, "A cluster-based approach to filtering spam under skewed class distributions," In HICSS, pp.53-59, 2007.

[19] S. Needleman and C. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," J. Mol. Biol, Vol.48, No.3, pp.443-453, 1970.

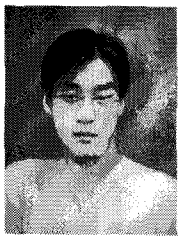
저 자 소 개

이 윤 정(Yun-Jung Lee) 정회원

- 1995년 2월 : 부경대학교 전자계산학과(이학사)
- 1999년 2월 : 부경대학교 전산정보학과(이학석사)
- 2008년 8월 : 부경대학교 전자계산학과(공학박사)
- 2008년 9월 ~ 현재 : 부산대학교 U-Port 정보기술사업단 박사후연구원

<관심분야> : 얼굴 애니메이션, 웹 컨텐츠 시각화

지 정 훈(Jeong-Hoon Ji) 정회원



- 2003년 : 경성대학교 컴퓨터공학(학사)
- 2005년 : 경성대학교 컴퓨터공학(석사)
- 2005년 ~ 현재 : 부산대학교 컴퓨터공학과 박사과정

<관심분야> : 프로그래밍언어 및 컴파일러, 프로그램 표절검사, 자바가상기계

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 조교수
- 2005년 ~ 현재 : 부산대학교 컴퓨터공학과 박사과정

<관심분야> : 프로그래밍언어 및 컴파일러, 함수형 언어, 그리드컴퓨팅, 소프트웨어 메트릭, 프로그램 시각화

조 환 규(Hwan-Gue Cho) 정회원

- 1884년 : 서울대학교 계산통계학과(학사)
- 1990년 : 한국과학기술원 전산학(공학박사)
- 1991년 ~ 현재 : 부산대학교 컴퓨터공학과 교수

<관심분야> : 알고리즘 이론, 생물정보학