인터넷 웹 게시판 분석을 통한 논쟁지수 계산모델 개발

박규태*, 송주원, 심정민, 조환규, 우균
부산대학교 컴퓨터공학과
{kelvinpark, hgcho, woogyun}@pusan.ac.kr,
lubnaldo@gmail.com, kildin@nate.com

Develeoping Argument-Index Computing Model, In Internet Discussion Bulletin Analysing

Kyu-Tae Park*, Ju-Won Song, Jung-Min Shim, Hwan-Gue Cho, Gyun Woo
Dept of Computer Science, Pusan University

요 약

최근 인터넷 게시판에서 사용자들의 토론과 논쟁이 큰 이슈가 되고 있다. 이러한 논쟁은 게시판을 읽는 사용자들에게 불필요 한 부분이거나, 혹은 관심을 보이는 부분이다. 만약 이러한 논쟁이 있는 구역을 게시판전체에서 논쟁지수를 계산하여 사용자에게 정보를 가져다준다면, 사용자가 직접 글을 파악하지 않고도 논쟁구역을 회피하거나 더 관심 있게 읽을 수 있다. 본 논문에서는 댓글 수나 조회 수 등의 개별적인 정보를 이용하여 파악하는 것이 아니라 글쓴이의 연속된 정보를 가지고 게시판의 특정 부분의 논쟁구역을 찾는 방법을 논하고자 한다. 그래서 논쟁이 있는 인터넷 웹 게시판 사용자들의 논쟁 밀도와 논쟁지수를 정의 하고, 이를 계산하는 방법에 대한 모델을 제시하고자 한다.

1. 서론

최근에 네티즌들의 논쟁이 벌어지는 게시판들이 많이 늘어났다. 예전의 게시판에는 거의 학술적인 내용으로 지식을 논쟁했었지만, 요즘은 네티즌들의 여론 참여도가 높아져서 큰 화제 거리가 일어나면 각종 시사 게시판들에는 엄청난 양의 글들이 쏟아져 올라오는 추세이다. 이렇게 쏟아지는 글들에는 조회수, 댓글수, 답글수 등을 기본적인 평가 가치로 매겨진다. 그 평가 가치를 기준으로 베스트글이나 메인글 등 핫이슈로 분류를 한다. 하지만 정작 조회수가 많은 글 혹은 댓글이 많은 글을 읽어 보면, 주제와는 전혀 다른 엉뚱한 내용이 있는 경우가 있으며, 의도적으로 조회수나 댓글을 조작하여 여러 용도로 후킹을 노리는 경우도 있다. 이러한 평가 가치는 한 개의 게시글에 대한 평가이며 많은 예외가 존재하기 때문에 평가 가치로서 부적절한 경우가 많다. 본 논문에서 소개하고자 하는 계산 모델은 게시판에 나열되어있는 일정 구역 다수의 글들에 대한 평가 값에 관한 연구이다.

특정 게시판에 나열되어 있는 글들은 어쩌면 연관성이 없는 독립적인 글일 수도 있다. 하지만 열 띈 논쟁이 이루어지면 한번 글을 썼던 사람이 얼마 지나지 않아 또 다시 글을 쓰게 된다. 논쟁이 과열되면 자주 답변성 글을 쓰게 되는 상황을 보면서, 특정 구역에 얼마나 뜨거운 논쟁이 벌어지는가를 계산 할 수 있을 것이라는 생각을 얻게 되어 본 연구를 시작하게 되었다.

먼저 특정 게시판의 게시물들의 HTML 소스를 가져와 원하는 정보만을 추출 한다[1]. 추출된 정보를 기반으로 살펴보면, 모든 게시글 하나에는 한명의 글쓴이가 존재한다. 하나의 게시글을 단위로 하여 그 글의 글쓴이가 몇 번째 만에 다시 글을 쓰게 되는가를 거리로 계산해, 각각의 글을 거리로써 계산 점수를 기록한다. 여기서 계산된 점수는 0을 기준으로 다시 쓴 글이 가까울수록 양수의 큰 값을, 멀수록 음수의 작은 값을 가지는데, 이 점수를 이용하여 특정 구간의 합이 가장 큰 값을 가지는 구간이 어느 곳인지 찾아낸다. 바로 그 구간이 논쟁이 뜨겁게 일어나고 있는 구간이다. 그다음 그 구간을 제외하고 다시 특정 구간의 합이 가장 큰 구간을 찾는 방식으로 계속 수행하면, 논쟁이 활발한 구간을 여러 부분 찾을 수 있다.

실험 대상으로는 국내의 열띤 토론이 일어나며 게시글의 수가 40만개 이상인 'MBC토론 게시판'과 하나의 게시글에 대하여 계속적인 답변글이 이어지는 형식을 띤 'SkepticalLeft 게시판'을 선정하였다. 이 두 게시판에서 글쓴이들의 순서로 하여금 논쟁 과열구간이 어느 곳인지 계산해보고 그 타당성을 검토하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 인터넷 웹 게시판 분석에 대한 관련연구를 살펴본다. 3장에서는 논쟁이 빈번한 게시판 데이터에 대한 선별을 하고, 4장에서는 수집한 데이터로 논쟁구역을 찾아내는 방법에 관하여 알아본다. 5장에서는 찾아낸 논쟁구역의 타당성을 검토하고 6장에서는 전체 실험에 대한 결과를 정리한다.

2. 관련 연구

한편 논쟁지수를 측정하는 다른 방법으로는 해당 글의 내용을 가지고 감정지수 등을 파악하는 방법이 있을 것이다. 비전문가의 글에서 키워드를 추출 하는 연구를 살펴보면, 약어 혹은 신조어 때문에 개별 글의 내용을 가지고 평가 가치를 측정하는 것에는 제약사항이 존재했다[2].

여러 블로그들 간의 상호작용 가시화 연구에는 블로그 간의 관계를 정의하고, 이를 가시화하여 보여준다[3,4]. 본 연구에서는 이를 표현할 수 있는 전 단계로서 논쟁구간을 정의하고, 논쟁구간의 가중치를 계산하는 모델을 연구하여, 추후에 가시화 할 수 있도록 선행하고자 한다.

"OLAP에서 MAX-of-SUM 질의의 효율적인 처리 기법"연구에서는 기가(GB)에서 테라(TB)에 이르는 방대한 양의 데이터베이스에서 의사 결정 지원 질의를 위한 영역 질의를 효과적으로 계산하기 위한 MAX-of-SUM 알고리즘을 이용하였다[5]. 계산된 논쟁지수 값들에서 과열된 논쟁구간을 검색해 내기 위한 도구로 MAX-of-SUM 알고리즘을 사용해보았다.

3. 웹 게시판의 논쟁 데이터 수집

웹 게시판 중에서 열띤 논쟁이 벌어지는 게시판이 요즘 많이 개설되고 있다. 그중에 활발한 논쟁이 벌어지는 MBC 토론게시판과 SkepticalLeft 게시판 두 가지를 실험 모델로 삼았다. 본 계산모델에서는 글쓴이의 순서를 가지고 논쟁지수를 계산하기 때문에 글쓴이가 나열 되어 있는 구조의 게시판을 실험데이터로 삼고자 한다.

MBC 토론게시판은 게시글과 댓글이 각각 다른 글로 이어지는 형식을 띈다. 그 게시판에서 40만개가 넘는 게시글의 HTML 소스를 받아와 해당 소스를 분석하여 데이터를 수집하였다. 수집한 데이터를 문서 분석기를 이용해 글쓴이들의 정보만 순서대로 받아와 논쟁지점과 논쟁지수를 파악할 수 있는 데이터를 수집하였다.

표 1 SkepticalLeft 게시판의 높은 논쟁 게시물

순번	제목	댓글 수	사람 수
1	엉뚱한 비교	146	16
2	스켑렙 과학자와 합리주의	131	18
3	김일성 이명박 히틀러	121	21
4	유승준	115	12
5	친 반미, 미 친한	109	16
6	철거민 저항	105	16
7	종부세 토론4	94	17
8	용산사태	81	12
9	불온서적의 추억	77	21
10	지젝과 라깡	75	10

SkepticalLeft 게시판은 MBC 토론게시판과 다르게 하나의 게시글안에 댓글이 계속 이어지는 형식으로 하나의 게시글을 하나의 게시판으로 취급하려 한다. [표 1]은 SkepticalLeft 게시판의 최근 1000개의 게시글을 수집하여 그 중 조회수와 댓글수가 가장 많은 글을 순서대로 10개 수집한 것이다. 그리고 하나의 글에서 댓글을 쓴 글쓴이들의 데이터를 수집하고 그 데이터를 분석하여 논쟁지점과 논쟁지수를 파악하였다.

4. 논쟁구역 검색

위 논쟁 게시물에서 논쟁 지점을 찾는 방식은 몇 가지 절차를 가진다. 첫째, 해당 게시물의 주소로 접근하여, html 소스를 가져온다. 둘째, 해당 소스를 분석하여 글쓴이들을 가져온다. 셋째, 글쓴이들의 순서를 파악하여 거리를 계산한다. 넷째, 거리를 기준으로 각 글쓴이들의 논쟁지수를 계산한다. 끝으로 계산된 논쟁지수로 MAX-of-SUM 알고리즘을 이용하여 논쟁 지점들을 검색해 나간다.

이러한 논쟁구역을 검색함에 있어 먼저 논쟁지수와 논쟁밀도를 정의해 보고자한다.

4.1. 논쟁밀도와 논쟁지수의 정의

논쟁밀도란, 특정 게시판의 특정 구역 내에 게시글의 수에 비례한 글을 쓴 사람 수를 나타낸다. 가령 게시글이 100개이고, 그 100개 게시물의 글쓴이가 100명이면 100/100이 되어 논쟁 밀도는 1이 된다. 논쟁 밀도 1이라함은 논쟁이 전혀 이루어지고 있지 않음을 시사한다. 이번엔 100개의 게시물 중에 글쓴이가 2명이면 2/100이 되어 논쟁 밀도는 0.02가 되어, 아주 왕성한 논쟁이 일어나고 있음을 시사한다.

논쟁지수란, 자신의 글에 대한 반박글이 있을 수 있고 이러한 반박글에 대하여 얼마나 가깝게 자신의 답변을 하게 되는가를 수치화 해놓은 것이다. 예를 들어 a라는 글쓴이가 하나의 글을 단위로 하여 글을 썼을 때 몇 번째 만에 다시 글을 쓰게 되는가를 거리로 계산한 값을 'd'로 하였다. 만약 글쓴이가 글을 쓰고 바로 다음에 또 쓰는 경우에는 논쟁으로 포함하지 않았다. 그리고 많은 실험을 통하여 글쓴이가 글을 쓸 때 8번째 이상으로 쓴 글의 경우는 논쟁적 평가 값이 낮다고 판단하였다. 그래서 1이하를 밑으로 하는 지수함수를 사용하여 지수가 7이면 0.5에 가깝도록 0.91을 지수함수의 밑으로 결정하였다. 함수식에서 0과 1사이의 값을 갖는 논쟁점수 값에서 논쟁구역 상수값 0.51을 빼주기로 하여 논쟁지수가 높으면 양수 값을 낮으면 음수 값을 갖도록 결정했다. 이를 수식으로 표현하면 다음과 같다.

$$S_d = R^{d-1} - C$$

S_d : 논쟁 점수　　d : 반박 거리

R : 논쟁구역 함수 의 밑　　C : 논쟁구역 상수

수식 1 논쟁지수 계산모델

여러 경우를 예로, 거리에 따라 논쟁지수가 높은 경우와 낮은 경우를 살펴보자. [표 2]는 [그림 1]에서 구한 거리를 가지고 글쓴이별로 논쟁지수를 구한 결과이다. 거리가 짧은 경우 논쟁지수가 높게 나오고 거리가 먼 경우 논

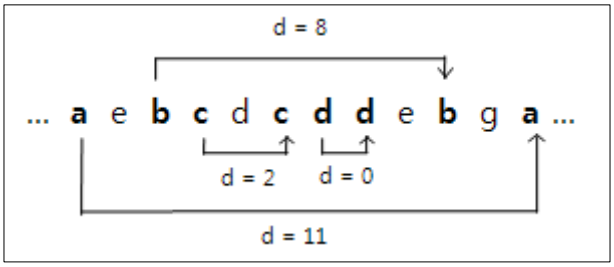

그림 1 거리 d 계산

표 2 논쟁지수 결과 값

글쓴이	거리	공식	논쟁지수
c	2	$S_d = 0.91^{2-1} - 0.51$	0.4
b	8	$S_d = 0.91^{8-1} - 0.51$	0.0067
a	10	$S_d = 0.91^{11-1} - 0.51$	-0.12

쟁지수가 낮게 나왔다. 논쟁지수가 높게 나온 부분의 글들이 정말 논쟁이 일어나고 있는지 확인하기 위하여 두 게시판에서 수집한 데이터들을 토대로 직접 확인해 본 결과 논쟁이 일어나고 있음을 확인하였다.

4.2. 논쟁구역 검색 방법

논쟁구역은 MAX-of-SUM 알고리즘을 이용하여 논쟁지수의 합이 가장 큰 구간이 논쟁이 활발히 이루어지고 있다고 판단한다. MBC 토론게시판 40만개의 게시물에서 각각 논쟁지수를 구하고 이를 MAX-of-SUM 알고리즘으로 검색해본 결과 [표 3]와 같이 두 명의 글쓴이가 서로 의견을 주고받으며 논쟁을 벌인 것을 직접 확인하였다.

표 3 논쟁지수 합이 가장 높게 나온 구역

게시물 제목	이 름	거리	논쟁지수
방송다시 한번 찬찬히...	오소현	2	0.0972
만족 하실런지 모르지만...	김근식	2	0.0972
Re: 만족 하실런지 모르...	오소현	2	0.0972
그건 진행자를 탓하심이...	김근식	2	0.0972
핫! 무슨말씀이신지??	오소현	2	0.0972
Re: 핫! 무슨말씀이신지??	김근식	2	0.0972
그럼...	오소현	2	0.0972
Re: Re: 핫! 무슨말씀...	김근식	2	0.0972
Re: Re: Re: 핫! 무슨말...	오소현	2	0.0972
Re: Re: Re: Re: Re: 핫!...	김근식	2	0.0972
Re: Re: Re: Re: Re: Re:...	오소현	2	0.0972
Re: Re: Re: Re: Re: Re:...	김근식	2	0.0972
어이없네	오소현	2	0.0972
Re: 어이없네	김근식	2	0.0972
그게아니구요	오소현	2	0.0972
앗 ! 아직도 불만이	김근식	6	-0.189
ㅡㅡ;;;;;;	오소현	54	-0.8061

현재 보여 지는 결과의 게시물 개수는 17개로 적절한 구간이 검색되었지만, 이와 같은 논쟁지수 계산식으로 다른 게시판 글들을 검색해보니 논쟁구역이 많게는 200개의 게시물이 결과로 나왔다. 원인을 찾아보았더니 논쟁이 높은 구간임에는 분명했지만, 구간 중간에 논쟁과 관련 없는 글들이 있음을 알게 되었다.

[그림 2]는 MBC 토론게시판에서 [수식 1]에서 논쟁구역 상수인 C값을 조절하여 얻게 된 논쟁구역의 길이를 나타낸 것이다. 0.35 이하일때부터 급격하게 넓은 논쟁구역이 검색되었으며, 해당 구역의 글을 읽어 보았을 때 논쟁의 연관성이 높은 결과 값을 얻은 C 값은 0.8128 이었다.

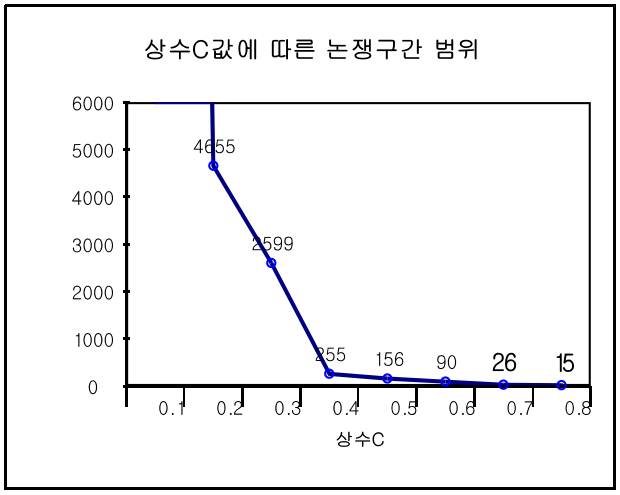

그림 2 상수C값에 따른 논쟁구간 범위

이는 논쟁구역 상수인 C값을 높게 하면 좁은 논쟁구역이 검색되고 낮게 하면 넓은 논쟁구역이 검색됨을 의미한다. 일반적으로 논쟁 밀도가 높은 게시판일수록 전반적으로 논쟁지수가 높게 계산되고, 이에 따라 논쟁구역이 넓게 계산되므로, 게시판의 성향에 따라 상수C를 변경해줄 필요가 있다고 판단했다.

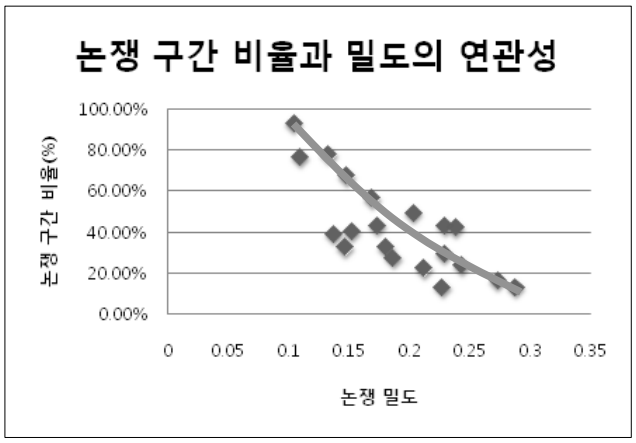

그림 3 논쟁밀도에 따른 논쟁구간의 크기 비율

논쟁구간이 100%라 함은 대상 게시판의 모든 글이 논쟁구간으로 검색되었음을 의미하는데, 논쟁밀도에 따른 MAX-of-SUM 계산 결과 값을 살펴보면 논쟁 밀도가 높

을수록 즉, 논쟁밀도 값이 작게 나올 수록 고정된 논쟁구간 상수 C값에 대해 대체적으로 넓은 논쟁구간이 검색되었다.

하지만 논쟁밀도와 논쟁구간의 상관관계에 대한 결과를 얻지는 못하여, 이 부분은 향후 과제로 남겨두기로 하였다. 현재 논쟁지수 계산모델에 따른 논쟁 구역을 검색할 때에는 사용자가 검색하고자하는 게시판의 논쟁밀도에 따라 적절한 논쟁구역 상수 C값을 결정하여 만족한 결과를 얻으면, 그 이후로 논쟁지수 계산식을 이용하여 논쟁구역들을 검색해낼 수 있다.

5. 평가 및 분석

논쟁지수 계산 모델로 여러 게시판의 논쟁구역을 검색해본 결과 논쟁이 과열된 구간을 찾을 수 있었으며, 찾아진 구간을 제외하고 지속적인 검색을 해보면 같은 게시판 내에 또 다른 논쟁구역을 찾을 수 있었다. 물론 점점 논쟁 과열도는 낮아진다. 이와 같이 게시판의 글쓴이의 순서를 가지고 논쟁구간을 검색 하는 방법은 빠른 속도로 과열된 논쟁구간을 찾을 수 있다는 장점이 있다. 현재 이용한 MAX-of-SUM 알고리즘은 선형-시간 복잡도를 가지고 있어, 40만건에서 100개의 구간을 검출해내는데 5초정도의 검색시간을 소요한다. 10개 미만의 구간에 대해서는 1초도 걸리지 않는 실시간성을 보장할 수 있다.

6. 결론

본 논문에서는 인터넷 게시판의 글쓴이 기반의 분석을 통해 논쟁 밀도와 논쟁지수를 파악하는 모델을 제안하였다. 이로써 논쟁이 벌어지는 특정 게시판에서 글쓴이의 순서만으로 논쟁이 과열된 구간을 검색해낼 수 있다.

이 모델로 실험 할 때에는 게시글이 만개가 넘는 큰 구역을 검색하는 경우 HTML 소스분석에 소요되는 시간만으로도 적게는 10분에서 하루 내도록 걸리는 경우도 있었다. 만약 알아보고자 하는 게시판의 게시글 데이터베이스를 보유하고 있다면 HTML 소스분석 없이 실시간으로 논쟁구역을 파악할 수 있다. 게시판을 운영하는 측면에서 본 논문에서 제시된 방법을 이용하여 사용자들에게 논쟁구역 확인 서비스를 제공한다면, 사용자들은 이러한 논쟁구역을 스스로 회피하거나 선택적으로 살펴볼 수 있는 장점을 지닌다[7].

실험 과정에서 논쟁이 과열되는 지역은 욕설이나 비하 발언이 많은 것을 조사하였다. 이로써 논쟁구역을 찾아내어 이 구역에 대한 집중적인 게시판 관리가 이루어진다면, 기존의 방식보다 더 효율적으로 관리할 수 있을 것이다[8].

향후 이러한 방법을 이용하여 논쟁 과열 구역을 컬러링을 하는 가시화를 하여 사용자들이 더욱 쉽게 논쟁구간을 파악하도록 연구를 해보거나, 게시판 전체의 논쟁 과열도를 구하는 등의 연구를 행할 생각이다. 그리고 RSS를 이용하여 게시판의 논쟁구역 형성도를 살펴보면, 논쟁이 일어나는 시간대 등을 찾아내는 방법도 연구해 볼 수 있을 것이다[9].

참고문헌

[1] 이미란, 조동섭, “웹 페이지 분석을 위한 Web - Picker 설계 및 구현,” 한국정보과학회 학술발표논문집, pp. 603~605, 2003

[2] K Imura, A Hiramtsu and K Nose, “Extraction of Relationship between Keywords Written by An Individual for Retrieval from BBS on Web,” Jido Seigyo Rengo Koenkai Koen Ronbunshu, 2003

[3] Indratmo, Julita Vassileva and Carl Gutwin, “Exploring blog archives with interactive visualization,” AVI '08: Proceedings of the working conference on Advanced visual interfaces, pp. 39~46, 2008

[4] Meishan Hu, Aixin Sun and Ee-Peng Lim, “Comments-oriented blog summarization by sentence extraction,” CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 901~904, 2007

[5] 정희정, 김동욱, 김종수, 이윤준, 김명호, “OLAP에서 MAX-of-SUM 질의의 효율적인 처리 기법,” 정보과학회논문지 데이터베이스, 제27권 제2호 pp. 165~174, 2000

[6] 김현영, 이동훈, 이지형, “군집화를 이용한 개인화된 스팸 댓글 필터링 시스템,” 한국지능시스템학회 추계학술대회 학술발표논문집, 제18권 제2호, pp. 84~87, 2008

[7] Judith B. Pena-Shaff, Craig Nicholls, “Analyzing student interactions and meaning construction in computer bulletin board discussions,” Computers & Education, Volume 43 Issue 3 Page 313, 2004

[8] J Zhang, EM Rasmussen, “Developing a new similarity measure from two different perspectives,” Information Processing & Management, Vol 37 Issue 2 pp. 279-294, 2001

[9] 양단, 김양훈, 김국보, “RSS를 이용한 블로그 검색 웹 로봇 설계,” 한국인터넷정보학회 추계학술발표대회, 제9권 제2호 pp. 343 ~ 347, 2008