http://dx.doi.org/10.5392/JKCA.2011.11.10.019

웹게시판에서 가상온도를 이용한 게시글의 인기 예측

Predicting the Popularity of Post Articles with Virtual Temperature in Web Bulletin

김수도*, 김소라**, 조환규**
부산대학교 U-Port정보기술산학공동사업단*, 부산대학교 컴퓨터공학과**

Su-Do Kim(kimsd@pusan.ac.kr)*, So-Ra Kim(srkim_11@pusan.ac.kr)**,
Hwan-Gue Cho(hgcho@pusan.ac.kr)**

요약

블로그는 사용자에게 자신의 의견을 표현하고 다른 사람들의 의견을 수렴할 수 있는 자유로운 의사표현 네트워크를 제공한다. 어떤 글은 사회적, 정치적 이슈를 몰고 다니기도 하며 또 어떤 글은 사용자의 관심을 끌지 못하고 지나가기도 한다. 글이 작성된 초기에 향후 얼마나 인기를 얻을지 예측한다는 것은 글의 저자, 블로거, 광고회사 그리고 웹호스팅 모두에게 흥미로울 것이다. 인기를 예측하기 위한 다양한 연구들이 진행되어 왔지만 대부분의 연구들이 사용자간의 상호연관성에 기반하고 있고 정확한 값으로 표현하는데 높은 에러율을 발생하고 있다. 본 논문에서는 블로그에 글이 작성된 초기에 향후 글의 인기를 예측하기 위해 조회수를 사용하여 글의 인기를 4타입(explosion, hot, warm, cold)의 가상 온도로 예측하는 방법을 제안한다. 먼저 글의 포화시점을 정의하고, 초기 조회수와 포화시점 조회수의 관계를 통해 포화시점 조회수를 예측하는 모델링 공식을 유도하였다. 예측된 포화시점 조회수를 이용하여 글의 인기를 4타입의 가상 온도로 표현하였다. 초기 관찰기간에 따라 예측 정확률이 결정되고 있다. 실험결과 30분 이후부터 MAPE(Mean Absolute Percentage Error)가 30%이하로 낮아졌지만, explosive 타입의 경우 초기 조회수로 예측하기 힘들었다. explosive를 제외한 hot, warm, cold 타입에서는 30분후부터 86%이상의 평균 예측 정확률을 보여주며, 70분후부터는 90%이상의 평균 예측 정확률을 보여주고 있었다.

■ 중심어 : | 예측 | 인기 | 웹 블로그 | 소셜 네트워크 |

Abstract

A Blog provides commentary, news, or content on a particular subject. The important part of many blogs is interactive format. Sometimes, there is a heated debate on a topic and any article becomes a political or sociological issue. In this paper, we proposed a method to predict the popularity of an article in advance. First, we used hit count as a factor to predict the popularity of an article. We defined the saturation point and derived a model to predict the hit count of the saturation point by a correlation coefficient of the early hit count and hit count of the saturation point. Finally, we predicted the virtual temperature of an article using 4 types(explosive, hot, warm, cold). We can predict the virtual temperature of Internet discussion articles using the hit count of the saturation point with more than 70% accuracy, exploiting only the first 30 minutes' hit count. In the hot, warm, and cold categories, we can predict more than 86% accuracy from 30 minutes' hit count and more than 90% accuracy from 70 minutes' hit count.

■ keyword : | Prediction | Popularity | Web Blog | Social Network |

* 이 논문은 2010년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2010-371-B00008)
접수번호 : #110701-013
접수일자 : 2011년 07월 13일
심사완료일 : 2011년 09월 07일
교신저자 : 조환규, e-mail : hgcho@pusan.ac.kr

I. 연구 동기

웹 2.0 은 사용자에게 인터넷을 통해 다양하고 많은 정보를 쉽게 얻을 수 있고 제공할 수 있는 환경을 제공하였고 그 결과 등장한 것이 블로그와 소셜 네트워크 등이다. 블로그는 뉴스, 의견, 토론, 대화 등이 뒤섞인 커뮤니케이션 도구이다[1][2]. 인터넷 공간에서 블로그는 자유롭게 자신의 의견을 표현하고 다른 사람들의 의견을 수렴하는 여러 분야에 대한 의견 토론장이 되고 있으며 새로운 여론 형성의 장이 되고 있다. 어떤 글은 사회적・정치적 이슈를 몰고 다니기도 하고 또 어떤 글은 사용자의 관심을 끌지 못하기도 한다[3][4].

이전부터 콘텐츠의 인기를 예측하기 위한 여러 연구들이 진행되어 왔다[5-7]. 글의 인기를 예측할 수 있다는 것은 사용자의 흥미를 일으키는 주제, 파워블로거 선정을 통한 광고마케팅 활용, 또는 잘못된 글을 통한 사회적 논쟁 유발을 사전에 선별 관리하는 등 저자, 블로거, 광고회사, 웹호스팅 회사 등 모두에게 흥미로운 주제일 것이다. 그러나 대부분의 연구들이 사용자의 관계 분석을 통한 소셜 네트워크에 기반하고 있어 블로그처럼 사용자의 관계가 명확하지 않은 경우 예측이 어렵고, 정확한 값으로 예측하는데 높은 에러율을 발생하고 있다[5]. 대부분의 토론 블로그는 주관적인 평가에 따라 인기가 높은 글을 분류하여 사용자가 쉽게 볼 수 있도록 제공하고 이런 분류는 사용자의 관심을 더 유도하기도 한다. 사용자는 인기가 낮은 글보다 인기가 높은 글에 더 많은 관심을 주는 경향이 있기 때문이다[6].

글의 인기를 표현하는 요소는 관점에 따라 조회수, 댓글수, 추천수, 투표수 등 다양하다. 보통 영화에서는 관객 수, 드라마에서는 시청자 수로 흥행을 결정하는 것처럼 블로그에서 글의 인기를 표현하는 가장 객관적이고 정량적인 수로 표현되는 것이 조회수이다. 즉 사용자로부터 높은 조회수를 받으면 인기 있는 글이고 낮은 조회수를 받으면 인기 없는 글로 설명될 수 있다.

본 논문에서는 블로그의 특성을 분석하고 글의 인기를 판별하는 요소중 조회수를 사용하여 글의 초기에 향후 글의 인기를 예측하기 위해 먼저, 조회수 분석을 통해 글의 포화시점을 정의하였다. 둘째, 초기 조회수를 이용하여 포화시점 조회수를 예측하는 모델을 정의하였다. 셋째, 예측모델을 통해 글의 인기를 4가지 타입(explosive, hot, warm, cold) 가상 온도로 표현하였다.

본 논문의 구성은 다음과 같다. 2장에서는 인기 예측에 관한 기존 연구들을 살펴보고, 3장에서 온라인 블로그의 특징을 비교·분석한다. 4장에서는 제안하는 글의 가상 온도를 예측하는 방법과 실험결과를 기술하고, 마지막으로 5장에서 결론을 맺는다.

II. 관련 연구

웹 블로그와 소셜 네트워크에 관한 여러 연구가 진행되고 있다[3][5][12][13]. 지금까지 네트워크의 분석을 통해 연결망 분석, 사용자 행동 패턴 분석과 콘텐츠 인기 예측에 대한 연구들이 진행되었고, 특히 유명한 소셜 네트워크 서비스인 DIGG사이트에서 많은 연구가 있었다[6][7][11]. DIGG사이트는 콘텐츠를 등록하고 공유하는 사용자들이 참여하고 투표하는 방식의 소셜 뉴스 웹사이트이다. 사용자가 특정 콘텐츠의 링크나 스토리를 등록하면 “Upcoming”에 보이고, 회원들의 많은 추천(Digg)를 받은 글은 ”Top News”로 승진되는 유명한 소셜 네트워크 사이트이다[17].

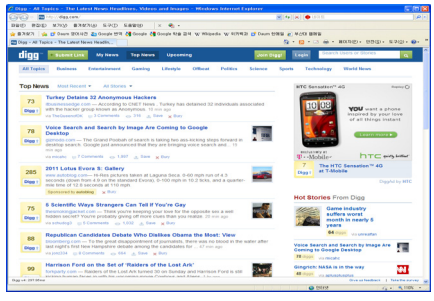


그림 1. DIGG사이트에서 “Top News” 첫화면

Lerman은 DIGG사이트에서 사용자의 행동분석을 통해 시간당 사용자수, 팬수, 페이지수, 스토리수 등 통계값을 계산하고, 통계값과 초기 투표수을 이용하여 스토리 인기를 계산한 후 최종 투표수를 예측하는 투표의 다이나믹성을 나타내는 수학적 모델을 제안하였다. 또한 콘텐츠의 추천, 필터링 그리고 평가에서 중요한 역

할을 하는 소셜 정보 처리에 관한 연구를 수행하였다[7-10]. Jamali와 Rangwala는 DIGG사이트에서 댓글 정보를 이용하여 스토리 인기를 예측하였다. 초기 시간 동안의 댓글수와 긍정적인 댓글인지 부정적인 댓글인지를 통해 Digg-score를 계산하고, 분류와 통계모델을 이용하여 콘텐츠의 인기를 예측하는 방법을 제안하였다[11]. Szabo와 Huberman는 DIGG와 YouTube사이트에서 콘텐츠가 제출된 후 초기 측정값(DIGG는 1시간후 투표수, YouTube는 7일후 조회수)과 30일 이후 측정값의 로그 변환을 통해 선형관계모델을 제안하고, 선형계수값과 노이즈값을 계산하여 인기를 예측하였다[6]. Lee, Moon 그리고 Salamatian은 dpreview.com과 myspace.com사이트에서 콘텐츠의 정확한 인기를 예측하기 힘들기 때문에 콘텐츠가 인기가 있을 것인지 가능성을 측정하는 방법론을 제안하였다. 생존기간, 댓글수, 조회수 등과 같은 관찰 가능한 객관적 요소를 이용하여 위험요소를 선택하고, Cox의 비례 위험 회귀 모형을 이용하여 생존기간과 댓글수 등을 예측하였다[5]. 대부분의 연구들이 사용자의 관계에 기반한 소셜 네트워크에 기반하고 있어 토론블로그처럼 사용자의 관계가 명확하지 않은 경우 적용하기 어렵고, 글의 인기에 영향을 미치는 여러 요소들로 인해 정확한 값으로 예측하는 데는 어려움이 있어 초기 높은 에러율을 발생하고 있다.

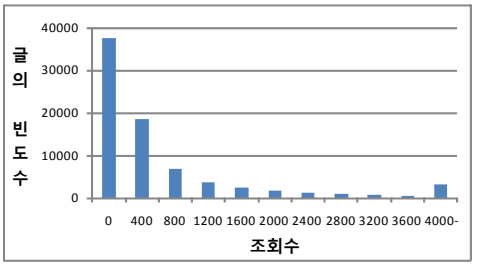
III. 온라인 블로그의 특성

본 장에서는 블로그의 특성을 분석하기 위해 온라인 정치 토론 사이트로 유명한 서프라이즈와 아고라에서 2010년 1월 1일부터 2010년 12월 31일까지 제출된 글을 조사하였다. [표 1]은 서프라이즈의 노짱토론방과 아고라의 자유토론방의 2010년 자료를 비교 및 분석한 표이다. 글의 수는 아고라가 많았지만 글의 평균 조회수는 서프라이즈가 상대적으로 높았다. 서프라이즈는 페이지당 110개의 글을 보여주고 있고, 아고라라는 20개의 글을 보여주고 있다[18][19]. 서프라이즈에서 글의 평균 조회수는 1043.54이며 아고라에서는 101.88로 조사되었다. 그러나 대부분의 글의 조회수는 평균보다 매우 낮아 표준 편차가 매우 높았다.

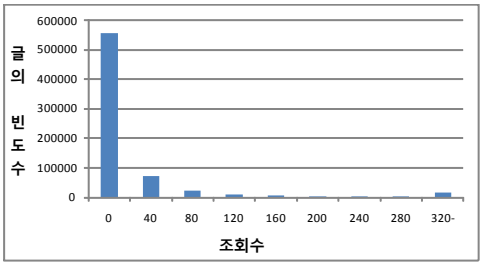
표 1. 2010년 서프라이즈와 아고라의 특성

2010년	서프라이즈	아고라
전체 글의 수	78,287	694,174
새로운 글의 수/일	214.48	1,901.85
조회수/글	1,043.54	101.88
최대 조회수	309,595	234,760
조회수 표준편차	4,203.10	1,314.91
저자수	21,790	33,343
게시글의 수/저자	3.59	20.82

[그림 2]는 글의 조회수에 따른 빈도수를 히스토그램으로 보여주고 있다. [그림 2](a)에서처럼 서프라이즈에서 조회수 1000을 넘는 글의 개수는 18,055(23.06%)이었다. [그림 2](b)에서처럼 아고라에서는 조회수 100을 넘는 글의 개수는 53,235(7.67%), 그리고 조회수 1000을 넘는 글의 개수는 7,579(1.09%)로 조사되었다. 조회수가 50이하인 글의 개수가 84.19%로 대부분의 글의 조회수가 매우 낮았다. 서프라이즈와 아고라에서 많은 글이 사용자로부터 높은 조회수를 받고 있었다.



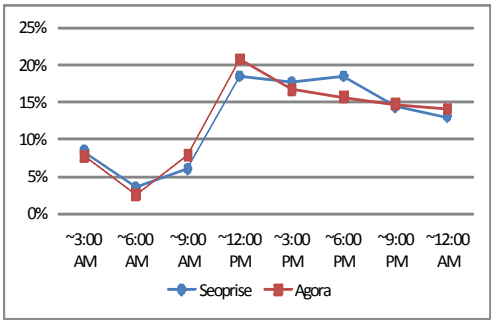
(a) 서프라이즈에서 조회수에 따른 글의 빈도수



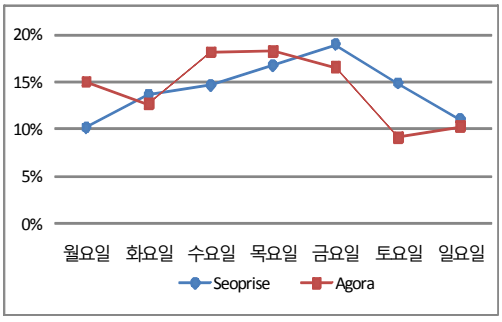
(b) 아고라에서 조회수에 따른 글의 빈도수

그림 2. 2010년 1월 1일에서 12월 31일까지 글의조회수에 따른 빈도수 히스토그램 [16]

[그림 3](a)는 시간에 따른 글의 작성율을 보여준다. 서프라이즈와 아고라 모두 오전 9시 이후 글의 수가 증가하고 있고 오후 6시 이후 감소하는 패턴을 보여주고 있다. 글의 수가 가장 많이 증가하는 시간은 오전 9시부터 오후 12시 사이였다. [그림 3](b)는 요일별 글의 작성 패턴을 보여주며, 서프라이즈와 아고라가 약간 달랐지만 모두 수요일에서 금요일사이 작성율이 높았고 주말에는 모두 감소하는 패턴을 보여주고 있다.



(a) 서프라이즈와 아고라에서 시간별 글의 작성율



(b) 서프라이즈와 아고라에서 요일별 글의 작성율

그림 3. 2010년 1월 1일에서 12월 31일까지 시간별 또는 요일별 글의 개수를 통한 작성 비율 [16]

[그림 4]은 1페이지에서의 조회수의 평균 변화량을 시간별로 표시하였다[16]. [그림 3](a)의 글의 작성율에서처럼 대부분의 사용자들이 오전 9시부터 오후 6시까지 가장 활동적으로 토론에 참여하고 있으며, 오전 6시에서 오전 9시까지인 출근시간대가 가장 낮은 참여율을 보여주고 있었다. [그림 5]는 페이지에 따른 평균 조회수 변화량을 보여주며, 페이지가 증가할 때마다 조회수가 급격하게 감소하고 있다.

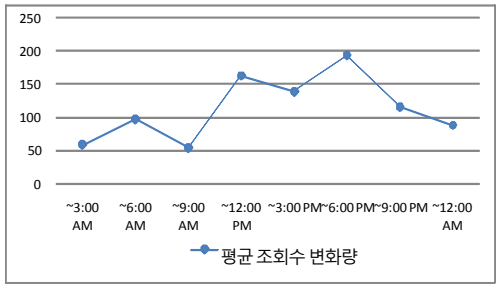


그림 4. 2011년 3월 15일~3월 18일 서프라이즈에 제출된 310개 글들의 1페이지에서 조회수 평균 변화량

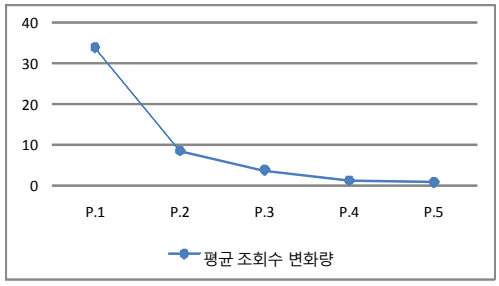


그림 5. 2011년 3월 15일~3월 18일 서프라이즈에 제출된 310개 글들의 페이지별 조회수 평균 변화량

IV. 인터넷 글의 인기도 예측 방법

본 장에서는 먼저 용어를 정의하고, 본 논문에서 제안하는 인기도 예측 모형과 글의 인기를 표현하기 위한 가상 온도에 대해 설명한다.

1. 용어 정의

본 절에서는 수식에 사용되는 기초 용어들을 정의한다. 먼저 대상 웹게시판 W에서 a_i는 전체 글의 집합 $A=\{a_o,a_1,a_2,\ldots,a_n\}$에서 i번째로 게시된 글을 말한다. 그리고 W에 게시된 전체 글 A를 관찰하기 위한 관찰시점들의 집합을 $T=\{t_o,t_1,t_2,\ldots,t_n\}$라고 할 때, $t_j<t_{j+1}$ 을 만족한다. 따라서 본 논문에서는 전체 글 A를 t_0부터 t_n까지 모두 $(n+1)$번 관찰하고, 관찰간격은 10분단위로 하였으며, 네트워크와 웹서버의 상황에 따라 약간의 오차가 있었다. 글 a_i의 j번째 관찰시점을 $time(a_i,j)$로 표시하며, $birth(a_i)$는 글 a_i가 W에

처음 게시된 시점, 즉 $birth(a_i)=time(a_i,0)$가 된다.

다음으로 본 논문에서 사용되는 글의 조회수와 관련된 용어들을 정의하고 설명한다. 관찰시점 t_j에서 글 a_i의 조회수를 $hit(a_i,t_j)$로 표시하며, $hit(a_i,t_j) \le hit(a_i,t_{j+1})$ 을 만족한다. 그리고 W에 게시된 $\forall a_i$의 $time(a_i,j)$에서의 조회수 집합을 $H_j=\{hit(a_i,time(a_i,j)) : \forall a_i \in A\}$로 표시한다. $\Delta hit(a_i,t_j)$는 관찰시점 t_j에서 글 a_i의 조회수 변화량을 나타내며, 식 (2)로 계산된다.

$$\Delta hit(a_i,t_j)=hit(a_i,t_j)-hit(a_i,t_{j-1}) \quad (2)$$

본 논문에서는 글 a_i가 W에서 활동이 정지된 시점을 $freeze(a_i)$로 정의하고, 24시간동안 조회수 평균 변화량이 임의의 작은 값 ε보다 작은 시점으로 식 (3)으로 표현한다. [그림 4]처럼 시간에 따른 사용자의 참여율이 매우 다르기 때문에 24시간동안의 조회수 평균 변화량을 이용하였다. d는 24시간을 나타내는 144, ε는 1/144로 지정하였다.

$$freeze(a_i)=t_n \ , \ \sum_{k=n-d}^{n} \Delta hit(a_i,t_k)/d < \varepsilon \quad (3)$$

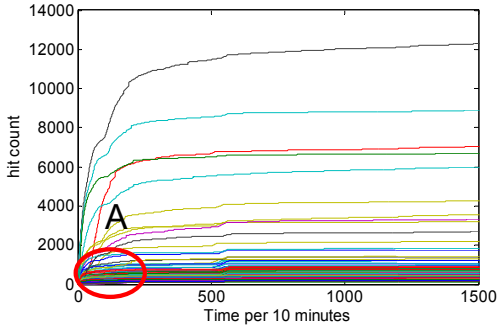
$varavg(a_i)$는 글 a_i의 $birth(a_i)$에서 $freeze(a_i)$ 전까지 조회수의 평균 변화량을 나타내며, 만약 $birth(a_i)=t_m$이고 $freeze(a_i)=t_n$이라고 할 때, $varavg(a_i)$는 식 (4)로 표현된다.

$$varavg(a_i)=\frac{1}{n-m}\sum_{k=m}^{n-1}\Delta hit(a_i,t_k) \quad (4)$$

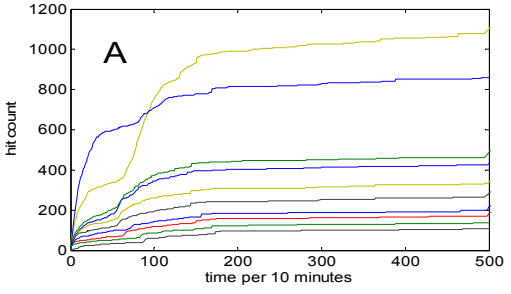
2. 실험 데이터

실험 데이터를 위해 사용자 의견교환과 공유가 활발히 이루어지고 있는 토론 블로그들 중에서 서프라이즈(Seoprise)의 노짱토론방에서 매 10분 간격으로 데이터를 수집하였다. 서프라이즈는 유명한 정치 포탈 블로그로서 하루 평균 230개의 글이 작성되며, 2010년 조회수가 5000이 넘는 글이 3000개 이상인 사용자 참여성이 높은 블로그중 하나이다[18]. 데이터를 분석하기 위한 훈련 데이터(training data)와 예측을 평가하기 위한 테스트 데이터(test data)으로 구분하였다. 훈련 데이터는 2011년 3월 15일부터 3월 20일까지 제출된 816개의 글을 추적하여 생성하였고, 테스트 데이터는 2011년 5월 5일부터 5월 10일까지 제출된 1157개의 글을 추적하여 생성하였다.

[그림 6]은 훈련 데이터에서 200개 글의 처음 게시된 시점인 $birth(a_i)$부터 10분간격으로 추적한 조회수를 표시한 그래프이다. 어떤 글은 사용자로부터 인기가 높아 조회수가 빠르게 증가하고 있어 위쪽으로 그래프가 그려지고 있지만, 대부분의 글들은 낮은 조회수로 그래프가 아래쪽으로 그려지고 있다.



(a) 200개 글들의 관찰시간에 따른 조회수 변화 그래프



(b) (a)의 A부분 글들의 조회수 변화 그래프

그림 6. 2011년 3월 15일~3월 20일 서프라이즈에 제출된 200개 글들의 $birth(a_i,0)$부터 10분간격으로 관찰된 조회수를 표시한 그래프

3. 예상 포화시점의 예측

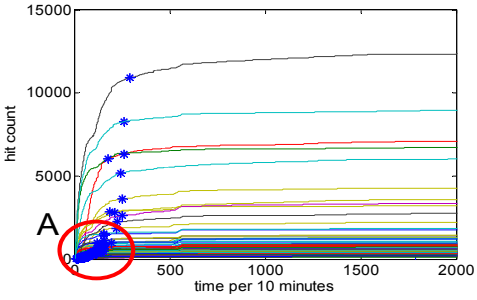
글의 작성된 후 글의 인기에 영향을 주는 요소들은 다양하다 : 조회수, 댓글, 페이지, 작성시간, 요일, 저자, 제목 등. 영화나 TV에서 인기를 표현하는 가장 중요한

요소는 관객수 또는 시청자수이다. 블로그에서 글의 인기를 반영하는 요소는 관점에 따라 사용자에 따라 다양할 수 있지만 가장 객관적이고 정량적인 값으로 표현되는 것이 조회수이다. 본 논문에서는 글의 인기를 결정하는 요소로 조회수를 사용한다.

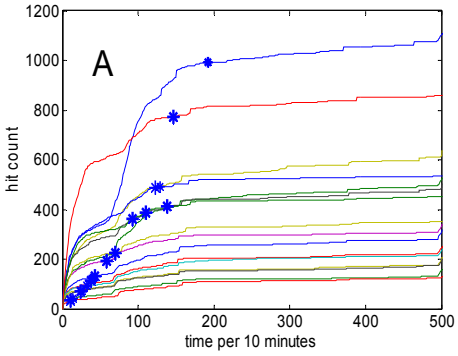
영화의 경우 관객수는 영화가 개봉된 직후부터 급격하게 증가하다가 일정한 시간이 지나면 서서히 감소되어 마침내 상영이 종영된다[14]. [그림 6]에서 글의 조회수도 영화에서처럼 작성된 초기에 급격히 증가하다가 어느 시점이 되면 서서히 감소되는 경향을 보이고 어느 시점부터는 거의 증가되지 않는 형태를 보여주고 있다. 영화에서의 경우 관객수가 급격하게 증가하다가 감소되기 시작하는 시점까지의 관객수가 영화의 흥행을 거의 결정하는 것처럼 글의 조회수도 초기 급격하게 증가하다가 감소되기 시작하는 시점까지의 조회수가 전체 조회수의 평균 88.45%를 차지하고 있다.

본 논문에서는 이 시점을 "포화시점"으로 정의하고, 글 a_i의 포화시점 $sat(a_i)$는 어떤 시점 t_j에서의 조회수가 $hit(a_i,t_j)$일 때, 시점 t_j가 되기전 24시간동안의 조회수 평균 변화량이 $varavg(a_i)$보다 작은 시점들 중에서 가장 빠른 시점으로 정의하고, 식 (5)로 나타낸다. 시간에 따른 사용자의 참여율이 매우 다르기 때문에 24시간동안의 조회수 변화량의 평균값을 사용하였다. [그림 7]은 식 (5)에 의해 계산된 글의 포화시점을 표시한 그림이다.

$$sat(a_i) = \arg min\{t_j \mid t_j \in T\}, \quad \frac{\sum_{k=j-d}^{j} \Delta hit(a_i,t_k)}{d} < varavg(a_i) \tag{5}$$



(a) 200개 글들의 예상 포화시점(*)



(b) (a)의 A부분 글들의 예상 포화시점(*)

그림 7. 서프라이즈에 제출된 200개 글의 식 (5)에 의해 계산된 포화시점(*) 표시

글의 작성된 후 예측을 실행하는 초기 시점은 빠르면 빠를수록 효과적이지만 충분하지 않은 데이터로 예측시 많은 오차를 발생시킬 것이고, 너무 늦은 시점에서 예측시 효과는 매우 낮을 것이다. 예측을 실행할 측정시점(t_j)을 구하기 위해 MATLAB의 피어슨 상관계수 함수를 이용하여 글의 초기 조회수 집합 H_j와 포화시점 조회수 집합 H_{sat}의 상관관계를 조사하였다. [표 2]는 MATLAB에서 구해진 글의 시간별 조회수 집합와 포화시점 조회수 집합의 상관계수 r을 표시하였다.

[표 2]에서 30분이후 조회수와 포화시점 조회수가 0.8 이상의 강한 상관관계를 가지고 있었다. 측정시점(t_j)을 30분이후로 지정할 수 있으며, 측정시점 조회수를 이용하여 포화시점 조회수를 예측한다. 포화시점 조회수는 정지시점 조회수와 0.9966이상의 아주 강한 상관관계를 가지고 있어 포화시점 조회수를 통해 글의 인기를 결정한다.

표 2. 초기 조회수 집합(H_j)과 포화시점 조회수 집합(H_{sat})의 피어슨 상관계수 r값. 글이 게시된 후 30분부터, 즉 $time(a_i,3)$부터 포화시점 조회수와 강한 상관관계가 있음을 보여주고 있다.

r \ j	1	2	3	4	5	6	7
$r(H_j, H_{sat})$	0.59	0.79	0.81	0.82	0.83	0.90	0.92

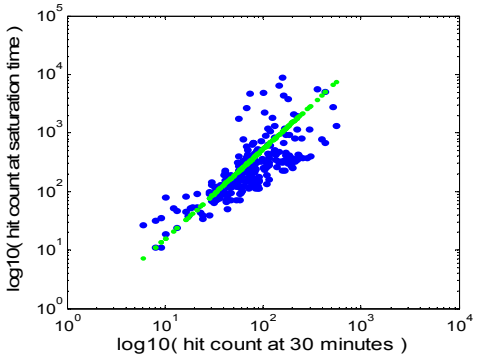
4. 인기도 예측 모형

[그림 8]은 MATLAB에서 로그변환을 통해 초기 조회수 집합을 x축, 포화시점 조회수 집합을 y축으로 표시한 그래프이다. 초기 조회수와 포화시점 조회수의 로그 변환 그래프를 통해 두 조회수가 서로 선형관계에 있다는 것을 알 수 있으며, 이를 통해 초기 조회수를 이용하여 포화시점 조회수를 예측하는 선형 모델링 공식을 유도하였다. [그림 8](a)는 30분 조회수와 포화시점 조회수의 관계를 보여주고, (b)는 60분 조회수와 포화시점 조회수의 관계를 보여주고 있고, (c)는 120분 조회수와 포화시점 조회수의 관계를 보여주고 있고, 그리고 (d)는 300분 조회수와 포화시점 조회수의 관계를 보여주고 있다. 초기 조회수와 포화시점 조회수의 로그변환 값이 시간이 지날수록 점점더 선형모델로 밀접하게 모여드는 것을 알 수 있다.

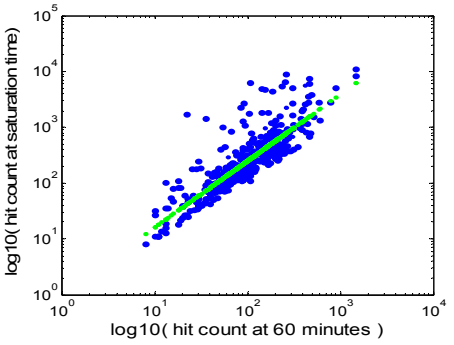
글의 초기 조회수와 포화시점 조회수가 로그변환후 선형관계를 가진다는 것을 이용하여, 어떤 글 a_i의 포화시점의 예측 조회수 $predict(a_i|j)$는 글이 제출된 후 j번째 관찰시점에서의 조회수 $hit(a_i,time(a_i,j))$를 이용하여 예측 가능하며, 식 (6)으로 나타낸다. α와 β는 MATLAB의 Curve Fitting 툴을 이용하여 구하였고, $predict(a_i|j)$은 [그림 7](a), (b), (c) 그리고 (d)에서 초록색 선으로 표시하였다.

$$predict(a_i|j) = \alpha \bullet hit(a_i, time(a_i,j))^{\beta} \qquad (6)$$

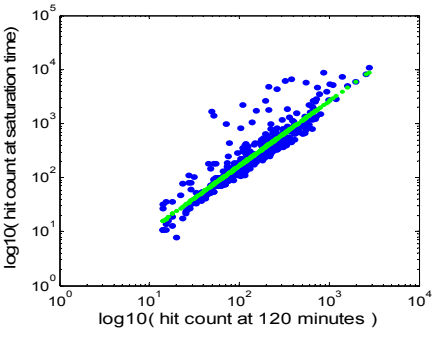
if $j=3$, $\alpha=0.47$ $\beta=1.53$



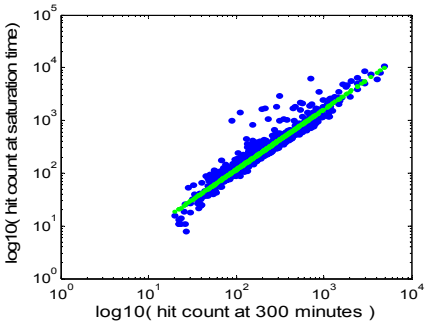
(a) 30분 조회수(H_3)와 포화시점 조회수(H_{sat})의 로그변환



(b) 60분 조회수(H_6)와 포화시점 조회수(H_{sat})의 로그변환



(c) 120분 조회수(H_{12})와 포화시점 조회수(H_{sat}) 로그변환



(d) 300분 조회수(H_{30})와 포화시점 조회수(H_{sat}) 로그변환

그림 8. 초기 조회수와 포화시점 조회수의 로그변환을 통한 선형관계 표시. x축은 log10(초기 조회수), y축은 log10(포화시점 조회수)를 나타내며, 초록색 선은 식 (6)에 의해 예측된 글의 포화시점 조회수

초기값을 이용하여 향후 글의 인기를 정확한 값으로 예측하는 것은 매우 힘들어 높은 오차율을 발생하고 있다[5]. 사용자는 글의 정확한 조회수보다 글의 인기가 높은지 낮은지에 더 영향을 받고 있다.

본 논문에서는 글이 인기를 표현하는 방법으로 포화시점의 예측 조회수를 이용하여 4 타입 가상 온도로 표현한다. 물질의 뜨겁고 찬 정도를 나타내는 온도 타입을 이용하여 폭발적 인기를 뜻하는 explosive타입, 높은 수준의 인기를 뜻하는 hot타입, 보통 수준의 인기를 warm 타입, 매우 낮은 인기를 뜻하는 cold타입으로 구분하여 표현하였다. 어떤 글 a_i의 포화시점에서의 실제 가상 온도를 $vtemp(a_i)$라고 할 때, $vtemp(a_i|j)$는 글의 j번째 관찰시점에서 예측한 $predict(a_i|j)$를 식 (7)에 의해 4가지 타입중 하나로 표현한 예측 가상 온도이다.

$$vtemp(a_i) = \{ \text{explosive, hot, warm, cold} \} \quad (7)$$

$$vtemp(a_i|j) \begin{cases} \text{explosive}, & 1500 \le predict(a_i|j) \\ \text{hot}, & 500 \le predict(a_i|j) < 1500 \\ \text{warm}, & 70 \le predict(a_i|j) < 500 \\ \text{cold}, & predict(a_i|j) < 70 \end{cases}$$

[그림 9]는 글의 실제 가상 온도와 예측 가상 온도를 비교한 예시 그림이다. 검은색 실선은 시간에 따라 변화되는 실제 조회수 변화 그래프를 나타내고, 빨간색 점선은 예측된 조회수 변화 그래프를 나타낸다. 그리고 검은색 실선 그래프에서 글의 실제 포화시점의 가상온도를 파란색 점으로 표시하였고, 빨간색 점선 그래프에서 예측된 포화시점의 가상온도를 빨간색 점으로 표시하였다.

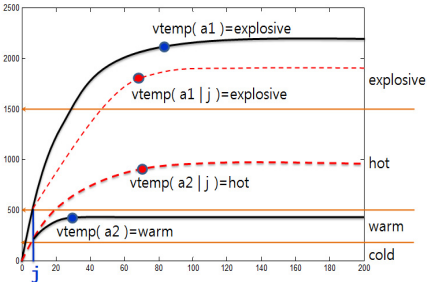


그림 9. 실제 가상온도(파란색 점)과 예측 가상온도(빨간색 점)의 예시. 검은 실선은 실제 조회수 변화 그래프고, 빨간색 점선은 초기값을 이용하여 예측한 조회수 변화 그래프를 표시하였다.

글 a1의 실제 포화시점의 가상온도 $vtemp(a_1)$=explosive이다. 글 a1의 j번째 관찰시점에서 예측한 포화시점 조회수 $predict(a_1|j)$=1800 이라고 할 때, $vtemp(a_1|j)$=explosive로서 실제 가상온도와 일치한다. 글 a2의 경우 실제 포화시점의 가상온도 $vtemp(a_2)$=warm이다. 그러나 글 a2의 j번째 관찰시점에서 예측한 포화시점 조회수 $predict(a_2|j)$=900 이라고 하면, 예측 가상온도 $vtemp(a_2|j)$=hot으로 실제 가상온도와 일치하지 않아 예측이 실패한다. 만약 관찰시점 j가 길어질수록, $vtemp(a_i) \cong vtemp(a_i|j)$ 되어 예측 정확률은 높아질 것이다.

인기를 구분하는 기준은 훈련 데이터의 조회수 분석을 통해 상위 20%를 hot, 하위 20%를 cold, 그리고 나머지는 warm 타입으로 구분하고, 최상위 5%를 explosive 타입으로 구분하였다.

5. 실험 및 결과

표 3. 4타입의 가상 온도로 표현한 시간에 따라 예측된 글의 수. 42,69,690,356은 측정된 실제 글의 수

시간 (10분단위)	E	H	W	C
실제값	42	69	690	356
t= 1	18	54	934	151
t= 2	6	49	861	241
t= 3	8	61	783	305
t= 4	7	63	786	301
t= 5	8	66	775	308
t= 6	10	62	771	314
t= 7	10	62	757	327
t= 8	11	64	744	338
t= 9	11	63	748	335
t=10	12	61	742	342
t=11	14	59	726	358
t=12	13	59	726	359

E : explosive, H : hot, W : warm, C : cold

[표 3]은 실험 결과를 보여주고 있다. 수치데이터 42, 69, 690, 356은 4타입의 가상 온도로 표현한 실제 글의 개수이고, 나머지는 시간에 따라 예측된 글의 개수를 보여주고 있다. hot, warm, cold 타입에서 시간이 지날수록 실제 수과 예측 수는 점차 줄어들고 있으며 30분부터 86%이상 맞추고 있다. 그러나 explosive타입에서 실제 수와 예측 수의 차이가 매우 컸다. explosive 타입

의 글들 중에서 어떤 이유로 인해 우리의 예상보다 더 많이 그리고 긴 시간동안 사용자의 인기를 끌고 있어 초기 조회수를 이용하여 포화시점 조회수를 예측하기가 매우 힘들었다. hot타입은 훈련 데이터의 상위 20%까지로 구분하였지만 테스트집합에서는 전체 글의 약 14%를 차지하고 있었고, cold타입은 훈련 데이터의 하위 20%까지의 조회수를 통해 구분하였지만 테스트집합에서는 전체 글의 30%를 초과하고 있었다. 즉 테스트집합의 글이 훈련 데이터의 글보다 사용자의 관심을 끌지 못해 인기가 낮았다는 것을 알 수 있다.

표 4. 4타입의 가상 온도로 표현한 시간에 따라 예측된 글의 수와 실제 글의 수의 절대 오차율(PE:Percentage Error) 결과

(단위:%)

시간 (10분단위)	E	H	W	C
t= 1	57.14	21.74	35.36	57.58
t= 2	85.71	28.99	24.78	32.30
t= 3	80.95	11.59	13.48	14.33
t= 4	83.33	8.70	13.91	15.45
t= 5	80.95	4.35	12.32	13.48
t= 6	76.19	10.14	11.74	11.80
t= 7	76.19	10.14	9.71	7.87
t= 8	73.81	7.25	7.83	5.06
t= 9	73.81	8.70	8.41	5.90
t=10	71.43	11.59	7.54	3.93
t=11	66.67	14.49	5.22	0.56
t=12	69.05	14.49	5.22	0.84

E : explosive, H : hot, W : warm, C : cold

[표 4]는 4타입의 가상 온도로 표현된 예측된 글의 수(F_t)와 실제 글의 수(Y)와의 차이에서 발생하는 오차율(PE: Percentage Error)을 보여주고 있다. 오차율 $PE = |Y - F_t| / Y \bullet 100$로 계산된다. MAPE(Mean Absolute Percentage Error)는 10분과 20분에서는 43%으로 높았지만, 30분 이후부터 30%이하로 떨어졌다. explosive타입은 초기 조회수를 통해 예측이 매우 힘들어 오차율이 매우 높았다. 그러나 explosive를 제외한 hot, warm, cold 타입에서의 오차율은 30분부터 모두 15%이하로 낮아졌고, 70분후부터 10%이하로 낮아지고 있어 90%이상의 예측 정확률을 보여주고 있다. 시간에 따라 조회수 변화량이 일정하지 않기 때문에 예측 오차율도 약간씩 변하고 있다. [표 5]는 제안한 예측 방법을 적용한 예를 보여주고 있다. 글이 작성된 후 60분 시점에서 예측한 글의 포화시점 가상온도 예를 보여주고 있다. 60분 시점에서 hot, warm, cold는 모두 예측률이 85%이상으로 높지만 explosive의 경우는 예측이 힘들었다.

표 5. 서프라이즈에서 글이 작성된후 60분시점에서 예측한 글의 포화시점 가상온도 예. 결과 ○는 실제 온도와 예측온도가 일치한 경우, ×는 실제온도와 예측온도가 불일치한 경우.

번호	제목	저자	날짜	H_{sat}	실제 온도	H_6	예측 온도	결과
47469	나는...	배달...	2011/05/09	222	W	93	W	○
47467	민주...	컴온...	2011/05/09	261	W	228	H	×
47466	유빠...	sim...	2011/05/09	125	W	71	W	○
47464	여긴...	지나...	2011/05/09	329	W	298	H	×
47463	난...	난보...	2011/05/09	905	H	668	E	×
47462	정당...	낚시...	2011/05/09	272	W	121	W	○
47461	주식...	증권...	2011/05/09	100	W	57	W	○
47460	수도...	펌생...	2011/05/09	215	W	165	W	○
47459	참여...	손오...	2011/05/09	138	W	66	W	○
47458	농협...	컴맹...	2011/05/09	586	H	416	H	○
47457	유시...	sim...	2011/05/09	82	W	51	W	○
47456	부처...	행복...	2011/05/09	29	C	10	C	○
47455	유시...	김석...	2011/05/09	1674	E	576	E	○
47454	집권...	들불...	2011/05/09	86	W	52	W	○
47453	유시...	본질...	2011/05/09	161	W	80	W	○
47452	박지...	fx...	2011/05/09	53	C	31	C	○
47451	반금...	라몰...	2011/05/09	195	W	95	W	○
47450	어린...	mon...	2011/05/09	26	C	18	C	○
47449	손학...	야홍...	2011/05/09	124	W	70	W	○
47448	40년...	거다...	2011/05/09	4595	E	98	W	×

E : explosive, H : hot, W : warm, C : cold

V. 결 론

토론블로그를 통해 사용자들은 자유로운 커뮤니케이션 환경을 가지게 되었으며, 어떤 글은 사회적 또는 정치적 이슈를 일으키기도 한다. 본 논문에서는 글이 작성된 초기에 향후 글의 인기를 예측하기 위해 아래와 같이 제안하였다.

- 글에 대한 사용자의 관심이 충분히 반영된 시점인 포화시점을 정의하고 측정하였다. 글이 작성된 초기 사용자의 관심으로 조회수가 급격히 증가하다가 포화시점 이후 감소되는 경향을 보여준다. 포화시점 조회수는 최종시점 조회수의 평균 88.45%를 차지하여 포화시점 조회수를 통해 글의 인기를 결정할 수 있다.
- 글의 초기 조회수와 향후 조회수의 관계를 분석하여 초기 조회수와 포화시점 조회수가 로그변환후 선형관계를 가진다는 것을 보여준다. 이를 통해 초기 조회수로 포화시점 조회수를 예측하는 모델링 공식을 유도하였다.
- 글의 인기가 높고 낮음을 표현하는 방법으로 4타입의 가상온도를 이용하여 폭발적인 수준의 인기를 뜻하는 explosive타입, 높은 수준의 인기를 뜻하는 hot타입, 보통 수준의 인기를 뜻하는 warm타입, 매우 낮은 인기를 뜻하는 cold타입으로 구분하여 글의 인기를 표현하였다.

실험 결과, 글이 게시된후 30분부터 MAPE 에러율은 30%이하로, 2시간이후부터 7%이하로 내려갔다. warm과 cold 타입에서는 글이 게시된 후 30분 이후에 86%이상의 평균 정확률을 보였고 70분부터 90%이상의 높은 정확률을 보이고 있다. 그러나 hot과 explosive 타입은 warm과 cold에 비해 예측시간이 늦었으며 특히 폭발적인 인기를 가지는 explosive 타입의 글은 초기 높은 에러율이 발생하여 초기 단계에서는 예측하기 힘들었다.

블로그에서 대부분의 글이 비슷한 변화 패턴을 보여줄 때, 어떤 글의 조회수는 초기 낮은 증가량 또는 높은 증가량을 보이다가 향후 전혀 다른 그래프로 그려지기도 하여 예측이 매우 힘들었다. 향후 연구로 글의 인기에 영향을 주는 요소들의 관계와 예측 정확률을 높일 수 있는 개선 알고리즘에 대해 연구하고자 한다.

참 고 문 헌

[1] C. L. Lin and H. Y. Kao, "Blog Popularity Mining Using Social Interconnection Anaysis," IEEE Computer Society, Vol.14, pp.41-49, 2010.

[2] N. Agarwal, H. Liu, L. Tang, and P. S. Yu, "Identifying the influential bloggers in a community," Proc. of WSDM, pp.207-218, 2008.

[3] M. Gőtz, J. Leskovec, M. McGlohon, and C. Faloutsos, "Modeling Blog Dynamics," Proc. of the ICWSM, pp.26-33, 2009.

[4] Y. J. Lee, J. H. Ji, G. Woo, and H. G. Cho, "Analysis and Visualization for Comment Messages of Internet Posts," Journal of the Korea Contents Association, Vol.9, No.7, pp.45-56, 2009.

[5] J. G. Lee, S. Moon, and K. Salamatian, "An Approach to Model and Predict the Popularity of Online Conntents with Explanatory Fact ors," Proc. of WI-IAT, Vol.1, pp.623-630, 2010.

[6] G. Szabo and B. A. Huberman, "Predicting the Popularity of Online Content," Communication of the ACM, Vol.53, No.8, pp.80-88, 2010.

[7] K. Lerman, "Social Information Processing in Social News Aggregation," IEEE Internet Computing:special issue on Social Search, Vol.11, No.6, pp.16-28, 2007.

[8] K. Lerman, "Social Networks and Social Information Filtering on Digg," Proc. of ICWSM, 2006.

[9] K. Lerman and A. Galstyan, "Analysis of Social Voting Patterns on Digg," Proc. of WOSN, pp.7-12, 2008.

[10] K. Lerman and T. Hogg, "Using a Model of Social Dynamics to Predict Popularity of News," Proc. of WWW, pp.621-630, 2010(4).

[11] S. Jamali and H. Rangwala, "Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis," Proc. of WISM, pp.32-38, 2009.

[12] V. Gȯmez, A. Kaltenbrunner, and V. Lȯpez,

"Statistical Analysis of the Social Network and Discussion Threads in Slashot," Proc. of WWW, pp.645-654, 2008.

[13] A. Kaltenbrunner, V. Gòmez, and V. Lòpez, "Description and prediction of slashdot act ivity," Proc. of LA_WEB, pp.57-66, 2007.

[14] S. A. Ahn and T. J. Kim, "Clustering by Life Cycle of Motion Picture," the Korean Journal of Advertising, Vol.65, pp.61-76, 2004.

[15] D. Salvatore, *Schaum's outline of theory and problems of microeconomic theory, 3rd ed.*, McGraw-Hill Professional, 1992.

[16] S. D. Kim, S. H. Kim, and H. G. Cho, "P redicting the Virtual Temperature of Web-Blog Articles as a Measurement Tool for Online Popularity," Proc. of CIT, 2011.

[17] http://digg.com

[18] http://www.seoprise.com/board/list.php?table =seoprise_13

[19] http://bbs1.agora.media.daum.net/gaia/do/ debate/list?bbsId=D003

저 자 소 개

김 수 도(Su-Do Kim) 정회원

- 1995년 : 부경대학교 전자계산학과(이학사)
- 2001년 : 부경대학교 교육대학원 전산교육전공(교육학석사)
- 2008년 8월 : 부경대학교 정보시스템협동과정(이학박사)
- 2011년 ~ 현재 : 부산대학교 U-Port정보기술산학공동사업단 박사후연구원

<관심분야> : 웹 콘텐츠, 웹 블로그, 소셜네트워크

김 소 라(So-Ra Kim) 준회원

- 2011년 : 부산대학교 정보컴퓨터공학부 졸업(공학사)
- 2011년 ~ 현재 : 부산대학교 컴퓨터공학 석사 과정

<관심분야> : Bioinformatics

조 환 규(Hwan-Gue Cho) 정회원

- 1984년 : 서울대학교 계산통계학과(이학사)
- 1986년 : KAIST 대학원 전산학과(공학석사)
- 1990년 : KAIST 대학원 전산학과(공학박사)
- 1990년 3월 ~ 현재 : 부산대학교 컴퓨터공학과 교수

<관심분야> : 계산이론, 생물정보학