http://dx.doi.org/10.3745/KIPSTD.2012.19D.1.113

# 인터넷 토론 게시판의 게시물 인기도 예측 모델

이 윤 정[†] · 정 인 준[††] · 우 균[†††]


## 요 약

오늘날 인터넷 사용자들은 유튜브(YouTube)와 같은 온라인 콘텐츠 공유 사이트를 통해 손쉽게 자신의 콘텐츠를 만들고 다른 사람들과 공유하고 있다. 그로 인해 하루에도 엄청난 양의 온라인 콘텐츠들이 쏟아지고 있다. 온라인 콘텐츠들의 홍수 속에서 어떤 콘텐츠가 향후에 인기가 있을 것인지를 예측하는 문제는 일반 이용자들이나 콘텐츠 공유 사이트 운영자들 모두가 관심을 가지는 문제이다. 본 논문에서는 인터넷 토론 게시판에 등록된 게시물들의 인기도를 예측하는 방법을 제안한다. 본 논문에서는 인터넷 토론 게시판에 등록된 게시물들의 인기도를 예측하기 위해 게시물의 조회수를 인기 척도로 간주하고 각 게시물의 조회수 변화량을 분석하였다. 게시물의 최종 조회수를 예측하기 위하여 관찰된 조회수 시계열 데이터를 이용하여 지수 함수를 기반으로 하는 조회수 증가 모델을 제안한다. 다음 아고라 게시판의 게시물을 대상으로 한 실험에서 전체 실험 게시물 중 약 90.7%인 20,532개의 게시물이 예측 오차가 10개 이하로 나타났다.


**키워드 : 소셜 미디어, 인터넷 토론 게시판, 온라인 콘텐츠, 인기 예측, 온라인 콘텐츠 다이나믹스**

# A Model to Predict Popularity of Internet Posts on Internet Forum Sites

Yun Jung Lee[†] · In Jun Jung[††] · Gyun Woo[†††]


## ABSTRACT

Today, Internet users can easily create and share the digital contents with others through various online content sharing services such as YouTube. So, many portal sites are flooded with lots of user created contents (UCC) in various media such as texts and videos. Estimating popularity of UCC is a crucial concern to both users and the site administrators. This paper proposes a method to predict the popularity of Internet articles, a kind of UCC, using the dynamics of the online contents themselves. To analyze the dynamics, we regarded the access counts of Internet posts as the popularity of them and analyzed the variation of the access counts. We derived a model to predict the popularity of a post represented by the time series of access counts, which is based on an exponential function. According to the experimental results, the difference between the actual access counts and the predicted ones is not more than 10 for 20,532 posts, which cover about 90.7% of the test set.


**Keywords : Social Media, Internet Discussion Board, Online Contents, Predicting Popularity, Dynamics of Online Contents**

## 1. 서 론

개방과 참여로 대표되는 오늘날의 인터넷 환경은 인터넷 이용자들의 온라인 콘텐츠 생산 및 소비 형태의 변화를 가져왔다. 전문 콘텐츠 제작자가 아닌 보통의 이용자들이 직접 제작하고 온라인으로 등록한 콘텐츠들을 UCC(user created contents) 또는 UGC(user generated contents)라고 한다. 전 세계적인 온라인 동영상 커뮤니티인 유튜브(YouTube)의 경우 하루 평균 약 65,000개 이상의 동영상이 등록되는 등 국내외의 여러 사이트나 웹 블로그 등에서 하루에도 엄청난 수의 콘텐츠들이 등록되고 있다.

수많은 콘텐츠 중에서 많은 사람들의 관심을 받은 콘텐츠는 소수에 불과하며 이러한 인기 콘텐츠는 다른 콘텐츠에 비해 높은 조회수나 추천수를 기록하기도 한다. 여러 연구를 통해 온라인 콘텐츠의 조회수나 댓글의 분포는 거듭제곱 법칙을 따르며 대중의 관심이 소수의 콘텐츠에 집중되어 있다고 알려져 있다[1-3]. 즉 대부분의 콘텐츠들은 대중의 관심을 얻지 못하지만 콘텐츠의 인기가 어떤 임계치를 넘을 경우 폭발적으로 증가함을 알 수 있다. 따라서 어떤 콘텐츠가 인기가 있을 것인지를 예측하는 것은 사이트 운영자나 사이트 이용자 모두에게 중요한 관심사라고 할 수 있다. 사이트 이용자들은 많은 콘텐츠 중에서 볼 만한 콘텐츠를 쉽게 찾을 수

※ 이 논문은 2011년도 정부재원(교육과학기술부 인문사회연구역량강화사업비)으로 한국학술진흥재단의 지원을 받아 연구되었음(KRF-2011-371-B00008).
† 정 회 원 : 부산대학교 U-Port 정보기술공동사업단 박사후연구원
†† 준 회 원 : 부산대학교 컴퓨터공학과 석사과정
††† 종신회원 : 부산대학교 컴퓨터공학과 부교수(교신저자)
논문접수 : 2011년 11월 29일
수 정 일 : 1차 2011년 12월 27일
심사완료 : 2011년 12월 30일

있으며, 사이트 운영자 또한 인기 콘텐츠를 미리 예측함으로써 해당 콘텐츠를 사이트 전면에 배치하는 등의 조치를 취함으로써 이용자들에게 편의를 제공할 수 있다.

그러나 이용자들이 콘텐츠를 선택하는 기준은 다분히 주관적이며 개인적인 것이라 어떤 콘텐츠가 인기가 있을 것인지는 예측하기 어려운 일이다. 콘텐츠의 인기를 결정하는 데는 콘텐츠의 품질, 다른 이용자들의 관심, 작가의 영향력 등 여러 요소들이 복합적으로 작용한다. Salganik의 연구에 따르면 콘텐츠의 품질이 콘텐츠의 인기에 미치는 영향은 그다지 크지 않으며 오히려 작가의 사회적 영향력이나 다른 사람들의 선택에 대한 정보가 더 큰 영향을 주는 것으로 나타났다[4]. 최근 온라인 콘텐츠의 인기도를 분석하거나 예측하기 위한 몇몇 연구들이 시도되었다. K. Lerman은 소셜 뉴스 웹 사이트인 'Digg'를 대상으로 게시글의 투표와 추천 방법을 제시하기 위한 수학적 모델을 제시하였고, 게시글 작성자의 사회적 네트워크의 영향력을 고려하여 해당 게시물의 투표수를 예측하였다[5]. 또한 Sazbo와 Huberman은 온라인 콘텐츠의 인기도를 선형 모델을 이용하여 모델링하였다[6]. 이처럼 온라인 콘텐츠의 소비에 대한 통계적 특성이나 예측에 대해 다양한 연구가 진행되고 있으나 각 사이트마다 콘텐츠 접근에 대한 다양한 특성이 존재하며 이에 대한 더 많은 연구가 필요한 실정이다.

본 논문에서는 국내 유명 토론 게시판인 다음 아고라(Daum Agora)를 대상으로 게시판에 등록된 게시물의 초기 조회수 변화 데이터를 바탕으로 최종 조회수를 예측하는 방법을 제안한다. 이를 위해서 우선 아고라 게시판에 등록되는 게시물들의 시간에 따른 조회수 변화량을 관찰하고 이 데이터의 통계 분석을 통해 게시물들의 수명, 조회수 변화 동역학과 같은 통계적인 특성을 조사한다. 다음으로 게시물의 조회수 변화에 가장 적합한 수학적 모델을 추정한 후 실제 게시물들의 초기 조회수 변화 데이터를 이용해 제안 모델의 예측 가능성을 검증한다.

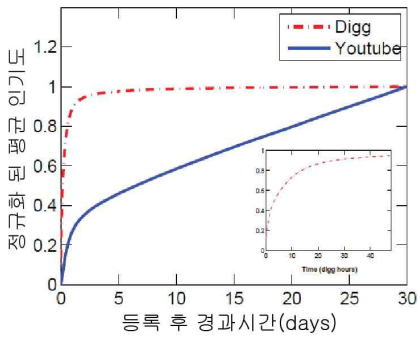
본 논문의 구성은 다음과 같다. 2장에서는 온라인 콘텐츠들의 통계적 특성 및 인기도 예측에 관한 기존 연구들을 살펴본다. 3장에서는 본 논문에서 사용한 데이터 집합과 통계적 특성에 대해 설명한다. 4장과 5장에서는 인터넷 게시물의 조회수 변화의 동적인 특성과 본 논문에서 제안하는 조회수 예측 방법에 대해 설명한다. 6장에서는 아고라 게시판의 게시물을 대상으로 한 실험을 통해 제안 방법의 예측 성능을 보인다. 마지막으로 7장에서 결론을 맺는다.

## 2. 관련 연구

최근 사회 연결망 서비스나 콘텐츠 공유 사이트의 이용이 활발해 짐에 따라 그에 관련된 연구들이 활발히 진행되고 있다. 이 장에서는 인터넷 뉴스, 게시물 또는 동영상과 같은 다양한 형태의 온라인 콘텐츠의 동적 특성과 인기 콘텐츠 예측에 관한 기존 연구들을 살펴본다.

K. Lerman은 뉴스 제공 사이트인 Digg에 등록된 뉴스 기사의 인기에 영향을 미치는 요소를 분석하고 뉴스 기사의 인기를 예측하는 수학적 모델을 제안하였다[7]. 이 연구에서는 Digg에서 뉴스 기사가 얻은 투표수를 그 기사의 인기로 간주하고 시간에 따른 투표수의 변화를 반영하는 모델을 제시하였다. 이 모델은 뉴스 기사가 받은 초기 투표수를 이용하여 해당 기사가 얼마나 흥미로운지를 평가하고, 작성자의 사회적 연결 정도를 함께 고려하여 향후의 투표수를 예측하였다.

Szabo 등은 콘텐츠 공유 사이트인 YouTube와 Digg에서 이용자들의 초기 접근 정보를 통해 향후의 인기도를 예측하는 방법을 제안하였다[6]. 그들은 콘텐츠의 조회수와 투표수의 증가량을 모델링함으로써 초기 데이터로부터 장기간의 변화 추세를 예측할 수 있음을 보였다. 실험에서 두 사이트의 콘텐츠 소비 패턴이 다르며 이러한 차이가 예측에 요구되는 초기 정보량의 차이를 가져오는 것으로 나타났다. (그림 1)은 두 사이트에서의 콘텐츠 인기도 변화 추세를 보여준다.




(그림 1) Digg과 YouTube 콘텐츠들의 인기도 변화 추세[6]

(그림 1)에서 Digg의 콘텐츠들이 YouTube 콘텐츠들 보다 인기도 증가 속도가 빠른 것으로 보아 Digg의 콘텐츠들의 이용자들의 관심을 받는 시간이 더 짧다고 할 수 있다.

Lee 등은 온라인 콘텐츠의 인기도가 아닌 인기 콘텐츠가 될 가능성을 추론하는 방법을 제안하였다[8]. 이 연구에서는 생존분석을 통하여 댓글수와 링크수 같은 객관적 관측 데이터에서 위험요소를 선택하고 이를 바탕으로 콘텐츠의 수명과 댓글수 등을 예측하였다. 두 개의 온라인 토론 사이트를 대상으로 한 실험에서 콘텐츠의 생존기간에 영향을 주는 5가지 위험요소로 댓글수, 콘텐츠 작성자가 쓴 댓글수, 댓글 작성자의 수, 댓글 간 평균 경과 시간, 그리고 댓글 간 경과 시간의 분산을 선택하였다.

Kim 등은 국내의 웹 게시판을 대상으로 게시물의 인기를 예측하는 방법을 제안하였다[9]. 이 방법에서는 게시글의 인기를 4타입(explosion, hot, warm, cold)의 가상 온도로 정의하고 초기 조회수와 포화시점 조회수와의 관계를 통하여 게시글의 향후 인기를 예측하였다.

이외에도 온라인 콘텐츠의 인기를 예측에 관한 여러 연구들이 있었으나, 앞서 살펴본 연구들과 마찬가지로 실험에 사용된 사이트들마다 적용 방법에 상당한 차이를 보이고 있다[10-12]. 또한 대부분의 문헌에서 개별 콘텐츠들에 대한 정확한 예측 결과를 밝히고 있지 않아 전체적으로 어느 정

도의 예측 성능을 보이는지 파악하기 어렵다. 웹 블로그 공간이나 사회 연결망 서비스 자체에 대한 연구는 많이 진행되어진 상태이나 그 안에서 만들어지고 소비되는 온라인 콘텐츠에 대한 연구는 아직 시작 단계라고 할 수 있으며 따라서 온라인 콘텐츠의 통계적 또는 동적 특성이나 인기에 영향을 주는 요소들에 대한 더 많은 연구가 필요한 실정이다.

## 3. 데이터 수집 및 통계

본 논문에서는 국내 유명 포털 사이트인 다음(Daum)에서 운영하는 토론 게시판인 '아고라' 게시판의 토론 게시물들을 대상으로 개별 게시물들의 최종 수렴되는 조회수를 예측하고자 한다[13]. 이를 위해서 아고라의 '자유토론'에 등록된 게시물들의 조회수를 10분 단위로 관찰하였다. <표 1>은 데이터 수집 기간 및 데이터 통계를 정리한 것이다.

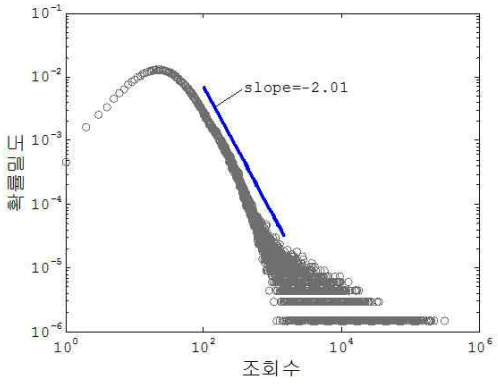
〈표 1〉 아고라 '자유토론' 게시판에서 수집된 게시물
- 데이터 수집 기간: 2011.03.30-2011.04.16

| 게시물 구분 | 개수 |
|---|---|
| 전체 | 25,931 |
| 1일 평균 | 1,440 |
| 2일 이상 관찰 | 22,639 |

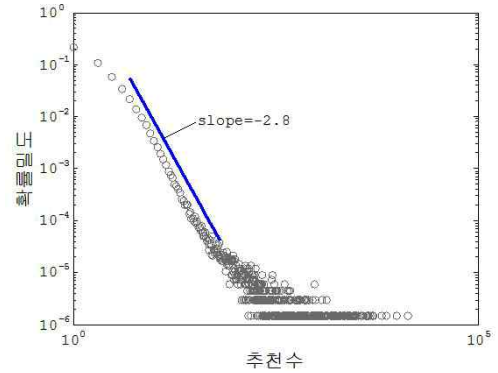
본 논문에서 게시물의 통계적 특성을 조사하기 위해 사용된 데이터 집합은 2011년 3월 29일부터 2011년 4월 15일까지 등록된 게시물로 구성되었다. 각 게시물에 대해 게시물 등록 시간부터 10분 간격으로 조회수, 추천 수, 댓글 수를 관찰하고 기록하였다. 관찰 기간 동안 등록된 전체 게시물의 수는 25,931개로 하루 평균 약 1,440개의 게시물이 등록되었다. 그 중에서 작성자에 의해 삭제되지 않고 2일 이상 관찰된 게시물의 수는 총 22,639개로 약 87.3%에 해당한다.

게시물의 인기는 주관적인 개념이라 정확히 정의되어 있지는 않지만 일반적으로 사용자들에게 많이 읽힌 게시물을 인기 게시물이라고 한다. 아고라 게시판에서는 게시물에 대한 이용자들의 반응을 조회수와 찬성, 반대, 그리고 댓글 수에 대한 정보로 제공하고 있으며 이러한 정보는 읽을 게시물을 선택하는 기준으로 작용한다. 게시물에 대한 이용자 반응 지표들의 분포는 (그림 2)와 같다.

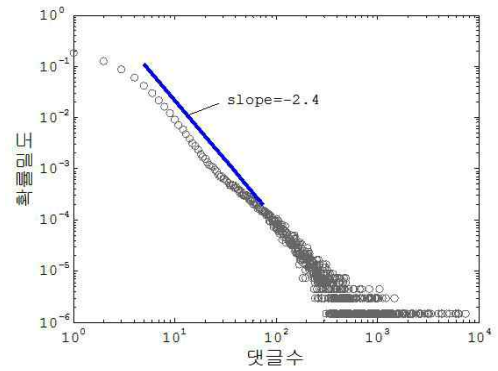
(그림 2)에서 추천 수는 찬성에 투표한 횟수를 나타낸다. 앞서 설명되었던 것과 같이 모든 지표들의 분포가 거듭제곱 분포를 따르는 것으로 나타났다. 각 그래프에서 기울기는 거듭제곱 분포의 지수를 의미한다. 이와 같이 게시물에 대한 이용자들의 반응을 그 게시물의 인기로 간주한다면 대부분의 게시물들은 많은 관심을 끌지 못하지만 일부 인기 게시물들은 상당이 많은 관심을 받는다는 것을 알 수 있다. 이 중에서 댓글 수와 추천 수의 경우는 한 건도 기록하지 않은 게시물들도 상당히 많고 조회수에 비해 상대적으로 적어 많은 실험 데이터를 얻기 힘들다. 따라서 본 논문에서는 게시물의 인기도를 반영하는 요소로 조회수를 선택하였다.



(a) 조회수 분포



(b) 추천 수 분포



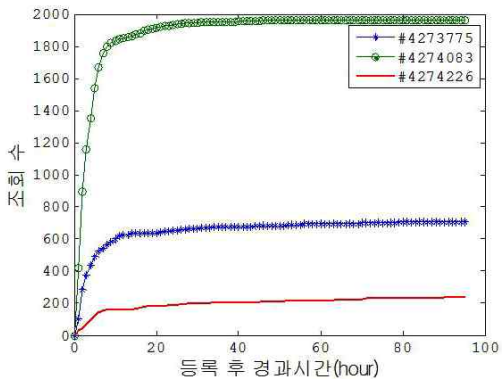
(c) 댓글 수 분포

(그림 2) 거듭제곱 분포를 따르는 게시물에 대한 이용자 반응의 분포 그래프

## 4. 조회수 시계열 분석

일반적으로 목록 형태로 되어 있는 게시판의 경우 새로 등록된 게시물이 목록의 상단에 위치하고 시간이 지날수록 다음 페이지로 넘어가게 되므로 게시물에 대한 접근성이 떨어진다고 할 수 있다. 본 논문에서는 등록 후 경과 시간에 따라 조회수 변화량이 어떻게 변하는지를 살펴보기 위해 일정 시간마다 게시물의 조회수를 관찰하여 시계열 데이터를 생성하였다. 아고라의 경우 가장 최근의 게시물이 목록의 상단에 게시되고, 한 페이지에 20개의 게시물이 보여진다. 마우

스 클릭을 통해 다음 페이지 목록으로 이동할 수 있다. <표 1>의 통계에서 하루 평균 약 1,440개의 게시물이 등록되므로 분당 1개씩의 게시물이 새로 등록된다고 할 수 있다. 즉 게시물이 등록된 후 약 20분 후면 다음 페이지로 넘어간다고 할 수 있다. 본 논문에서는 한 페이지 내에서의 위치는 게시물 접근성에 큰 차이가 없다고 간주하고 동일한 페이지에 있는 동안 두 번의 조회수 변화를 관찰하기 위해 조회수 변화 관찰 기간을 10분을 설정하였다. 본 논문의 데이터 집합은 전체 22,639개의 게시물에 대해 10분 단위의 조회수 시계열 데이터를 포함하고 있다. 그림 3은 아고라 게시물의 등록 후 경과시간에 따른 전형적인 조회수 변화 데이터를 보여준다.




(그림 3) 게시물 등록 후 경과 시간에 따른 조회수 변화 시계열

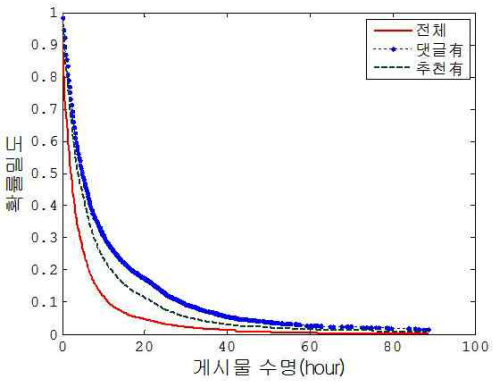
(그림 3)에 나타난 게시물의 정보는 <표 2>에 나타나 있다. 세 개의 게시물은 등록 시간도 서로 다르며 4일 이후의 조회수가 각각 705, 1,964, 236으로 상당한 차이가 있다. 그러나 세 게시물의 조회수 변화는 최종 조회수의 차이에도 불구하고 비슷한 패턴을 보임을 알 수 있다. 등록 후 처음 몇 시간 동안 조회수가 급격히 증가하다가 시간이 지날수록 더 이상 조회수가 증가하지 않고 최종 조회수에 수렴하는 형태를 나타낸다.

본 논문에서는 게시물이 등록 된 후 조회수가 일정 시간 내에 더 이상 증가하지 않을 때까지의 경과 시간을 게시물의 수명이라고 하고 생존 분석을 통해 게시물들의 수명을 조사하였다. 분석을 위해 6시간 동안 조회수가 증가하지 않으면 더 이상 조회수 변화가 없는 것으로 간주하였다.

본 논문에서 사용되는 게시물들의 생존 분석 결과는 (그림 4)와 같다. (그림 4)에서 각각의 생존 곡선은 게시물들의 수명 분포를 나타낸다. 분석 결과 대부분 게시물들의 수명이 하루를 넘지 못하는 것으로 나타났고, 댓글이나 추천을 받은 게시물들이 그렇지 않은 게시물들보다 일반적으로 수명이 더 긴 것으로 나타났다. 그렇다하더라도 대부분의 게시물들은 수명은 약 2일 이내인 것으로 나타나므로 본 논문에서는 게시물 등록 후 48시간 이후의 조회수를 해당 게시물의 최종 조회수라고 간주한다.

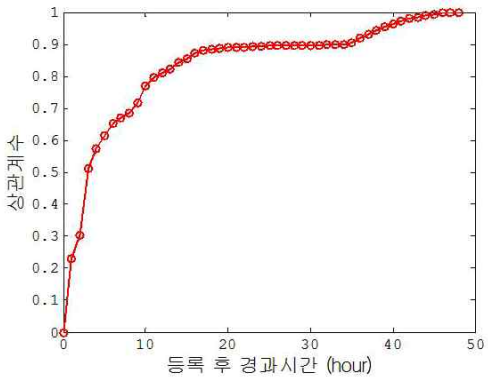
〈표 2〉 (그림 3)에 나타난 게시물 정보

| 게시물 번호 | 등록 시간 | 4일 후 조회수 |
|---|---|---|
| 4273775 | 11.03.29 12:35 | 705 |
| 4274083 | 11.03.29 17:25 | 1,964 |
| 4274226 | 11.03.29 19:19 | 236 |




(그림 4) 게시물의 생존 곡선

다음으로 게시물 등록 후 초반의 조회수 변화가 48시간 이후 조회수와 어떠한 관계가 있는지를 파악하기 위해 시간별 조회수와 48시간 이후 조회수의 상관관계를 조사하였다. (그림 5)는 시간에 따른 상관계수 변화를 보여준다. (그림 5)에서 상관계수 값은 피어슨 상관계수를 나타내며 시간이 지날수록 상관계수 값이 증가한다.




(그림 5) 48시간 이후 조회수와 경과 시간 별 조회수의 상관관계

조회수 관찰 기간이 길수록 더 정확한 예측을 할 수 있으나 관찰 기간이 너무 길면 대부분의 게시물이 조회수가 더 이상 증가하지 않으므로 예측의 의미가 없어진다. 따라서 본 논문에서는 상관계수가 0.5를 넘는 약 4시간까지를 초기 관찰 기간으로 정의한다. 이때의 상관계수 값은 약 0.57로 나타났다.

## 5. 조회수 변화 모델 및 예측

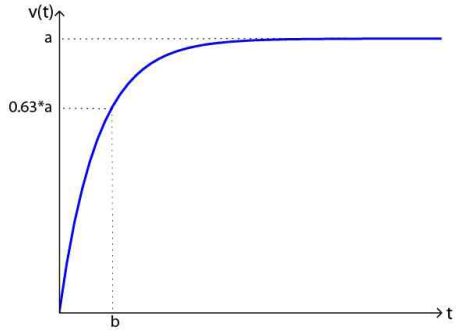
앞서 살펴본 바와 같이 게시물의 조회수는 등록 후 처음 몇 시간 동안 급격한 증가를 보이다가 시간이 지날수록 최

종 조회수에 수렴하는 형태를 보인다. 본 논문에서는 이러한 조회수 시계열 데이터를 식 1과 같은 지수 함수 형태로 모델링한다.

$$v(t) = a(1-e^{-t/b}) \quad (1)$$

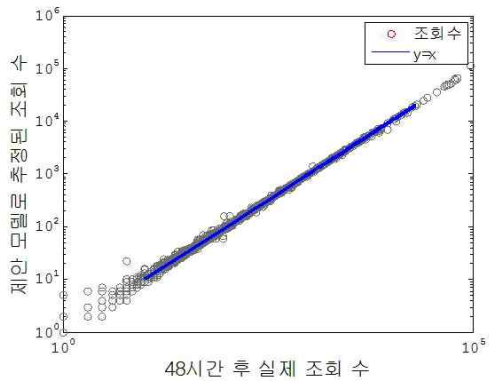
식 1에서 $t$는 등록 후 경과 시간을 가리키고, $v(t)$는 시간 $t$에서의 누적 조회수를 나타낸다.

(그림 6)은 식 1의 함수 그래프를 보여준다. (그림 6)에서 제안 모델의 전체적인 변화 추세가 (그림 3)에 나타난 조회수 변화 형태와 유사함을 알 수 있다. 또한 제안 모델에서 $a$는 최종 수렴될 $v(t)$의 값을 나타내고, $b$는 $v(t)$가 최종 수렴 값의 약 63%가 될 때의 $t$값을 나타낸다. 따라서 어떤 게시물의 조회수 시계열 데이터가 식 1로 피팅(fitting)되었다면 최종 조회수는 $a$이고 약 $b$ 시간 후에 최대 조회수의 약 63%를 얻을 것이라고 할 수 있다. 따라서 본 논문에서 제안한 조회수 변화 모델은 간단하지만 피팅된 함수에서 전체적인 조회수 변화 추이를 직관적으로 파악할 수 있다는 장점이 있다.




(그림 6) 조회수 변화 모델 그래프

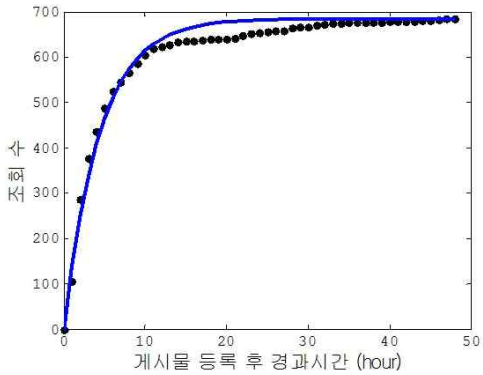
본 논문에서 제안한 조회수 변화 모델이 실제 조회수 시계열 데이터에 얼마나 적합한지를 알아보기 위해서 데이터 집합의 게시물들의 조회수 변화 데이터를 식 1의 함수에 피팅한 후 구해진 파라미터 값을 이용하여 48시간 이후의 조회수를 추정하였다. (그림 7)은 게시물 등록 후 48시간 후의 실제 조회수 관측 값과 추정된 값을 보여준다. (그림 7)에서




(그림 7) 실제 조회수와 제안 모델로 추정한 값과의 상관관계

가로축은 게시물 등록 후 48시간 지난 후의 조회수를 나타내고, 세로축은 제안 모델로 추정한 값을 나타낸다. 두 값들 사이의 상관계수는 0.997로 제안 모델이 실제 조회수 변화 패턴을 잘 반영함을 알 수 있다.

(그림 8)은 <표 2>에 설명된 4273775번 게시물의 조회수 관찰 데이터와 제안 모델로 추정한 함수의 그래프로 최종 조회수는 705회이고 48시간 지난 후의 조회수는 684이다. 제안한 함수 모델로 추정한 결과는 <표 3>과 같다.




(그림 8) 관찰된 조회수와 제안 모델로 추정된 함수 그래프

〈표 3〉 게시물 4273775에 대한 제안 모델의 조회수 추정 결과

| 48시간 후 조회수 | v(48) | a | b | $R^2$ |
|---|---|---|---|---|
| 684 | 685.2 | 685.2 | 4.415 | 0.9671 |

이 게시물에 대한 파라미터는 $a$가 685.2이고 $b$가 4.415로 추정되었다. 48시간 후의 조회수는 685.2회로 추정되어 실제 조회수와 1.2의 오차를 보였다. 또한 추정 함수와의 R2 값은 0.9671로 나타나 제안 모델이 실제 게시물의 조회수 변화 데이터를 잘 반영함을 알 수 있다. 따라서 게시물의 등록 후 처음 4시간 동안의 조회수 변화 데이터에 가장 적합한 식 1의 파라미터를 찾고, 이 값을 이용해 48시간 이후의 조회수를 예측할 수 있다.

## 6. 실험 및 결과

본 논문에서는 게시물의 초기 조회수 변화 데이터를 이용하여 최종 수렴될 조회수를 예측하는 방법을 제안하였다. 제안 방법의 예측 성능을 보이기 위해 <표 1>의 게시물들 중에서 등록 후 48시간 이상 관찰된 22,639개의 게시물을 대상으로 조회수 예측을 수행하였다. 실험은 초기 관찰 기간은 4시간으로 설정하였고, 10분 단위로 조회수를 관찰하였다. 예측 성능을 평가하기 위해 실제 관측치와 추정치의 절대오차($\epsilon_a$)와 상대오차($\epsilon_r$)를 식 2와 같이 각각 측정하였다.

$$\epsilon_a = |v_{est} - v_{obs}|, \quad \epsilon_r = \frac{|v_{est} - v_{obs}|}{v_{obs}} \quad (2)$$

식 2에서 $v_{obs}$는 게시물의 48시간 후 실제 조회수이고, $v_{est}$는 제안 방법으로 예측한 48시간 후의 예측 조회수 의미한다. 실험 게시물들의 예측 오차의 분포가 <표 4>에 나타나 있다.

〈표 4〉 제안 방법의 조회수 예측 오차 분포

| 절대 오차 | | | 상대 오차 | | |
|---|---|---|---|---|---|
| 오차 구간 | 게시물 수 | 비율 | 오차 구간 | 게시물 수 | 비율 |
| 10 이하 | 20,532 | 90.7% | 10%이하 | 10,481 | 46.3% |
| 20 이하 | 482 | 2.1% | 20%이하 | 3,628 | 16.0% |
| 30 이하 | 278 | 1.2% | 30%이하 | 2,444 | 10.8% |
| 40 이하 | 192 | 0.8% | 40%이하 | 2,415 | 10.7% |
| 50 이하 | 131 | 0.6% | 50%이하 | 1,594 | 7.0% |
| 60 이하 | 79 | 0.3% | 60%이하 | 441 | 1.9% |
| 70 이하 | 68 | 0.3% | 70%이하 | 459 | 2.0% |
| 80 이하 | 55 | 0.2% | 80%이하 | 246 | 1.1% |
| 90 이하 | 36 | 0.2% | 90%이하 | 153 | 0.7% |
| 100 이하 | 42 | 0.2% | 100%이하 | 209 | 0.9% |
| 100 초과 | 744 | 3.3% | 100%초과 | 569 | 2.5% |
| 합계 | 22,639 | 100.0% | 합계 | 22,639 | 100.0% |

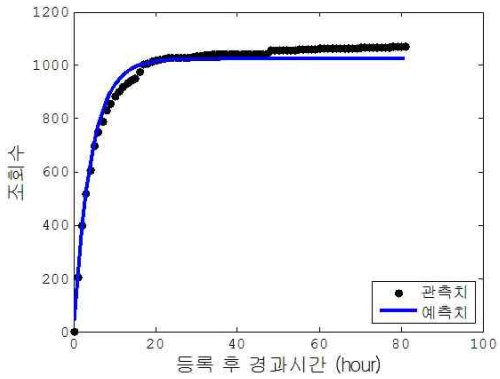
<표 4>에서 볼 수 있듯이 제안 방법으로 예측한 조회수의 절대 오차가 10 이하인 게시물이 전체의 약 90.7%를 차지한다. 적은 게시물의 경우는 절대 오차가 작더라도 상대 오차 값이 크므로, 조회수가 적은 게시물이 상대적으로 많은 실험 데이터의 경우도 상대 오차가 절대 오차보다 높게 나타났다.

<표 5>는 예측이 잘 된 게시물들 중 조회수가 높은 두 개의 게시물 정보를 보여준다. 두 게시물 모두 48시간 후의 조회수가 1,000회 이상으로 많은 이용자들의 관심을 받은 게시물이라고 할 수 있다. 게시물 등록 후 4시간 동안의 조회수 관찰 데이터를 이용하여 제안 모델로 48시간 후의 조회수를 예측한 결과 조회수 오차가 각각 32와 56으로 나타났다. 상대 오차는 3.0%와 5.5%로 산출되었는데, 이는 실제 조회수에 비해 오차가 적으므로 예측이 잘된 경우라고 판단할 수 있다.

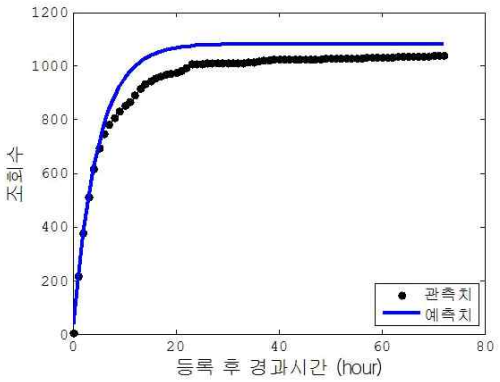
〈표 5〉 예측 오차가 적은 게시물의 추정 모델과 예측 결과

| 게시물 | 등록일 | 48시간 후 조회수 | | 예측오차 | | 추정 모델 | |
|---|---|---|---|---|---|---|---|
| | | 관측치 | 예측치 | 절대 | 상대 | a | b |
| 4287446 | 11.04.07 16:48 | 1,056 | 1,024 | 32 | 3.0% | 1024 | 26 |
| 4294047 | 11.04.13 14:51 | 1,027 | 1,083 | 56 | 5.5% | 1083 | 28 |

(그림 9)는 두 게시물이 전체 시계열과 추정 모델의 그래프를 보여준다. (그림 9)에서 두 게시물 모두 실제 관측치와 추정치 사이에 약간의 오차가 있으나 전반적으로 제안 모델이 전체 시계열을 잘 반영하고 있음을 알 수 있다.




(a) 게시물 4287445




(b) 게시물 4294047

(그림 9) 예측이 잘된 경우의 조회수 시계열과 추정 모델

<표 6>은 제안 모델로 예측이 실패한 경우의 게시물 예를 보여준다. 게시물 4283509는 48시간 이후 실제 조회수가 7,107인 것에 비해 예측 조회수는 17로 제안 모델이 실제 조회수 시계열을 반영하지 못한다고 추측해 볼 수 있다. 게시물 4280203의 경우도 상대오차가 95.2%에 달해 예측에 실패했음을 알 수 있다.

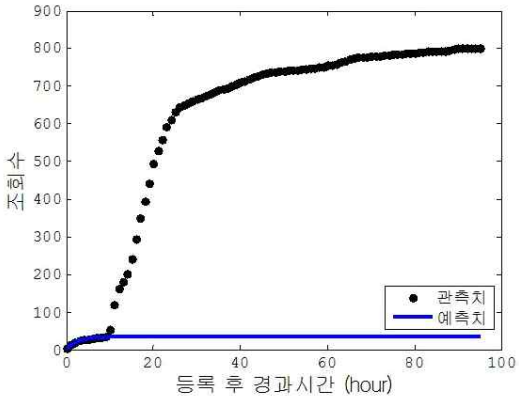
〈표 6〉 예측 실패한 게시물의 추정 모델과 예측 결과

| 게시물 | 등록일 | 48시간 후 조회수 | | 예측오차 | | 추정 모델 | |
|---|---|---|---|---|---|---|---|
| | | 관측치 | 예측치 | 절대 | 상대 | a | b |
| 4280203 | 11.04.03 00:50 | 739 | 35 | 704 | 95.2% | 35 | 13 |
| 4283509 | 11.04.03 14:30 | 7,107 | 17 | 7,079 | 99.7% | 17 | 2 |

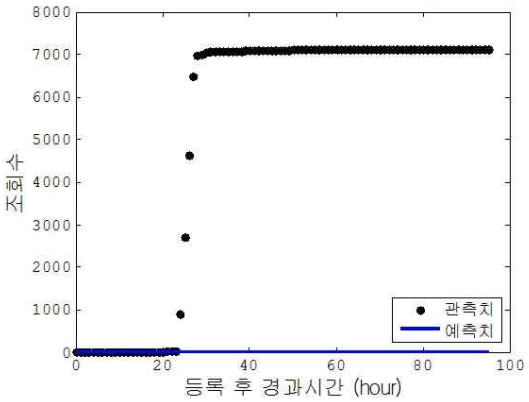
(그림 10)은 두 게시물의 전체 시계열과 추정 모델의 그래프를 보여준다. (그림 10(a))는 게시물 4280203의 조회수 변화 그래프로 등록 후 약 9시간 동안은 조회수 증가가 느리다가 그 이후에 급격히 증가하여 수렴하는 형태를 보인다. 이러한 원인으로 게시물 등록 시간이 사이트 이용자 수가 상대적으로 적은 새벽시간이라 이용자 반응이 늦은 것으로 추측해 볼 수 있다.

(그림 10(b))는 게시물 4283509의 조회수 변화 그래프로

앞서 살펴보았던 일반적인 게시물들의 조회수 증가 패턴과 완전히 다른 형태를 보인다. 게시물 등록 후 약 24시간 까지는 조회수가 30을 넘지 못하다가 그 이후 약 5시간 동안 약 7,000회의 조회수가 급격히 증가하였다. 이러한 패턴은 게시물 등록 시간의 영향이라기보다는 조회수 증가에 영향을 미치는 또 다른 요인이 있을 수 있음을 추측해 볼 수 있다. 실험 게시물들 중 조회수가 큰 게시물들 중에서 위의 두 게시물들과 같은 특이한 조회수 변화 패턴을 보이는 것들을 적지 않게 찾아 볼 수 있었다. 향후 이 게시물들에 대한 분석이 필요할 것이다.




(a) 게시물 4280203




(b) 게시물 4283509

(그림 10) 예측이 실패한 경우의 조회수 시계열과 추정 모델

## 7. 결론 및 향후 과제

본 논문에서는 국내의 인터넷 토론 게시판인 아고라를 대상으로 게시물들의 동적인 특성을 분석하고 향후 인기를 예측을 위한 연구를 수행하였다. 본 논문에서 기술한 연구 결과를 요약하면 다음과 같다. 첫째는 아고라에 등록된 게시물의 조회수 시계열 데이터를 이용하여 게시물들의 동적 특성을 분석하였다. 게시물의 등록 후 10분 단위로 관찰된 데이터에서 대부분의 게시물들의 조회수는 등록 후 빠른 증가를 보이다가 차츰 수렴되는 것으로 나타났다. 그리고 생존분석 결과 대부분의 게시물은 생존기간이 약 48시간을 넘지 않는 것으로 나타났으며, 초기 조회수와 수렴 조회수의 상관관계는 등록 후 4시간 이후부터 0.5이상으로 나타났다.

둘째로 게시물의 동적 특성을 기반으로 최종 수렴될 조회수를 예측하는 방법을 제안하였다. 게시물 조회수 증가 패턴은 경과 시간에 따른 지수함수 형태로 모델링 할 수 있었다. 이 모델은 간단하지만 추정된 계수를 통해 최종 조회수와 수렴 시기를 직관적으로 파악할 수 있도록 해준다. 게시물의 등록 후 4시간 동안의 조회수 변화 데이터를 이용하여 제안 모델을 추정하고 추정된 계수를 이용하여 최종 조회수 및 수렴 시간을 추정할 수 있었다.

마지막으로 제안 방법을 검증하기 위해 2011년 3월 30일부터 2011년 4월 16일까지 아고라의 자유토론방에 등록된 22,639개의 게시물에 대해 조회수를 관찰하고 제안 방법으로 최종 조회수를 예측하였다. 실험 결과 약 90.7%인 20,532개의 게시물들이 10 이하의 예측 오차를 보였다. 예측 오차가 큰 게시물들에서 등록된 후 오랜 시간 후에 급격한 조회수 증가가 발생하는 등의 다른 게시물들과는 다른 조회수 증가 패턴을 나타내는 특이한 경우를 볼 수 있었다. 이런 현상은 조회수가 높은 경우에 많이 발생하는 것으로 보였으며 제안 방법으로 예측하기 어려웠다.

제안 방법은 게시물의 등록 초기 조회수 변화를 이용하여 향후 최종 조회수와 수렴 시기를 예측할 수 있다. 따라서 인터넷 게시판이나 콘텐츠 공유 사이트에서 활용한다면 이용자들은 자신이 읽을 게시물을 선택하는데 도움이 될 수 있으며 사이트 운영자들은 인기 게시물 선정이나 사이트 배치 등이 이용할 수 있으며, 또한 제안 모델로 피팅되지 않는 게시물의 경우 특이한 조회수 증가 패턴을 보이는 것으로 추정할 수 있으며 의도적인 마우스 클릭에 의한 조회수 조작 등을 의심해 볼 수 있다.

본 논문에서 수행한 생존분석에 따르면, 댓글이나 추천이 있는 게시물이 그렇지 않은 게시물보다 수명이 더 긴 것으로 나타났다. 이는 댓글수와 추천수가 게시물의 생존기간이나 조회수에 미치는 영향이 크다는 것을 의미한다. 이러한 요소들의 영향을 예측 모델의 매개변수로 도입한다면 더 정확한 예측이 가능할 것으로 생각된다. 이는 향후 연구로 남겨둔다.

## 참 고 문 헌

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

[2] 이윤정, 지정훈, 우균, 조환규, “TRIB: 블로그 댓글 분류 및 시각화 시스템”, 정보처리학회논문지D, 제16-D권 제5호, pp.517-524, 2009.

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

[4] M.J. Salganik, P.S. Dodds, and D.J. Watts, “Experimental study of inequality and unpredictability in an artificial cultural

market," science, Vol.311, No.5762, pp.854-856, 2006, American Association for the Advancement of Science.
[5] K. Lerman, "Social Information Processing in News Aggregation," IEEE Internet Computing, pp.16-28, 2007.
[6] G. Szabo and B. A. Huberman, "Predicting the popularity of online content," Social Science Research Network Working Paper Series, November, 2008.
[7] K. Lerman and T. Hogg, "Using a Model of Social Dynamics to Predict Popularity of News," Proc. of WWW, pp.621-630, 2010(4).
[8] JG. Lee, S. Moon, and K. Salamatian, "An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors," in Proc. of WI-IAT 2010, 2010.
[9] 김수도, 김소라, 조환규, "웹게시판에서 가상온도를 이용한 게시글의 인기 예측", 제11권 제10호, pp.19-29, 2011.
[10] M. Tsagkias, W. Weerkamp, and M. de Rijke, "News Comments: Exploring, Modeling, and Online Prediction," in Proc. of the 7th ACD SIGCOMM conference on Internet Measurement, 2007.
[11] A. kaltenbrunner, V. Gomez, and V. Lopez, "Description and Prediction of Slashdot Activity," in LA-Web 2007, 2007.
[12] H. Yu, D. Zheng, B. Y. Zhao, and W. Zheng, "Understanding User Behavior in Large-Scale Video-on-Demand System," in Proc. of EuroSys'06, 2006.
[13] Daum 아고라, http://agora.media.daum.net/ (2011년 12월 방문)

**이 윤 정**

e-mail : leeyj01@pusan.ac.kr
1995년 2월 부경대학교 전자계산학과 (이학사)
1999년 2월 부경대학교 전산정보학과 (이학석사)
2008년 8월 부경대학교 전자계산학과 (공학박사)
2008년 9월～현 재 부산대학교 U-Port 정보기술공동사업단 박사후연구원
관심분야 : 컴퓨터 그래픽스, 웹 콘텐츠 시각화, 소셜 네트워크, 블로그 다이나믹스

**정 인 준**

e-mail : spd1335@gmail.com
2010년 8월 부경대학교 컴퓨터멀티미디어 공학부(학사)
2010년 9월～현 재 부산대학교 컴퓨터 공학과 석사과정
관심분야 : 이미지 프로세싱, 코딩 스타일 교육, 프로그램 시각화

**우 균**

e-mail : woogyun@pusan.ac.kr
1991년 한국과학기술원 전산학(학사)
1993년 한국과학기술원 전산학(석사)
2000년 한국과학기술원 전산학(박사)
2000년～2002년 동아대학교 컴퓨터공학과 전임강사
2002년～2004년 동아대학교 컴퓨터공학과 조교수
2004년～현 재 부산대학교 컴퓨터공학과 부교수
관심분야 : 프로그래밍언어 및 컴파일러, 함수형 언어, 그리드 컴퓨팅, 소프트웨어 메트릭, 프로그램 시각화