소셜 뉴스 집적 서비스에서의 카테고리별 뉴스 수명주기 패턴 분석

Pattern Analysis of News Lifecycle in a Social News Aggregation Service

원미경(Mikyoung Won)*, 이상진(Sangjin Lee)**, 이승준(Sungjun Lee)***,
박종헌(Jonghun Park)****

초 록

본 연구는 소셜 뉴스 집적 서비스(Social news aggregation service)에서 뉴스의 수명주기 패턴을 카테고리 별로 분석하여 사용자의 관심 변화를 예측할 수 있는 통계적 모델 제시를 목적으로 한다. 인터넷 뉴스는 사용자의 관심이 단시간에 집중되며 시간에 따른 사용자 관심의 쇠퇴가 명확하게 드러나는 웹 자원으로, 사용자 관심 변화에 대한 다양한 연구가 현재 진행 중에 있다. 본 연구는 뉴스의 수명주기를 카테고리 별로 분석하여 사용자 관심의 쇠퇴 정도를 예측할 수 있는 통계적 모델을 도출하였으며 소셜 뉴스 서비스 제공자(Social news aggregator)의 콘텐트 게시 정책이 사용자 관심의 급격한 성장을 발생시키는 주된 외부적 요인임을 분석하였다. 본 연구에서 제안된 인터넷 뉴스의 수명주기 모델은 독자의 관심을 지속시키면서 다양한 콘텐트를 공급하려는 소셜 뉴스 집적 서비스에 유용하게 적용될 수 있다.

ABSTRACT

The purpose of this paper is to present a statistical model that can predict the rapid shift of users' attention by analyzing the lifecycle patterns of news in a social news aggregation service. Internet news service sites have a distinct characteristic in a sense that users' attention change very quickly in a short period of time. In this research, we propose a regression model for each news category which can model the decay pattern of users' attention and the content promotion policy of a social news aggregator is proven to be a major source of the rapid growth in the popularity of news. The proposed model is expected to be useful for evaluation of the social news aggregation service provider's content promotion policy that attempts to maximize users' attention as well as the diversity of news contents.

키워드 : 소셜 뉴스 집적 서비스, 뉴스 수명주기, 패턴 분석
Social News Aggregation Service, News Lifecycle, Pattern Analysis

---

본 논문은 2007년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. R01-2007-000-11167-0).

* 다음 커뮤니케이션
** 교신저자, 서울대학교 산업공학과 박사과정
*** 서울대학교 산업공학과 석사과정
**** 서울대학교 산업공학과 부교수

2009년 01월 08일 접수, 2009년 04월 20일 심사완료 후 2009년 04월 30일 게재확정.

1. 서 론

최근 웹의 비약적인 발전과 함께 인터넷 뉴스가 새로운 미디어 매체로 성장하고 있으며[8], 인터넷 뉴스의 성장과 더불어 뉴스 콘텐트를 일반 사용자들에게 전달하는 소셜 뉴스 집적 서비스(Social news aggregation service)가 주목 받고 있다[16]. 소셜 뉴스 집적 서비스는 CNN[5]과 뉴욕 타임즈(The New York Times)[15] 등과 같은 다양한 뉴스 제공자(Publisher)로부터 전달되는 방대한 뉴스 콘텐트를 선별하여 사용자에게 전달하는 인터넷 서비스를 말한다. 이러한 소셜 뉴스 집적 서비스를 제공하는 대표적인 사이트로 현재 Digg.com[6], Reddit.com[19] 등이 있다.

소셜 뉴스 집적 서비스는 웹 페이지, 블로그 포스트(Blog posts) 등의 웹 콘텐트를 제공하는 다른 일반적인 웹 서비스와 마찬가지로 사용자 관심(Attention)의 극대화를 추구한다. 이는 기본적으로 사이트 방문수 및 광고 수입이 사용자의 관심에 비례할 뿐 아니라, 뉴스 콘텐트에 대한 평가 및 확산이 사용자들의 직접적인 참여에 의해 이루어지기 때문이다[21]. 이러한 소셜 뉴스 집적 서비스에서 사용자의 관심을 극대화하기 위해 기존에 다른 웹 서비스를 대상으로 연구된 사용자 관심 향상 알고리즘을 적용해볼 수 있다. 사용자 관심 향상을 위한 기존의 방법은 각 웹 콘텐트 별로 측정된 사용자 관심 값을 이용하여 웹 페이지 링크 순서 및 콘텐트 나열 순서를 재정렬한다[9].

반면, 소셜 웹 서비스 사이트에서 다루는 뉴스 콘텐트는 기존 연구에서 대상으로 하는 다른 웹 콘텐트와 구분되는 수명주기(Lifecycle) 상의 특성을 지니고 있다. 즉, 특정 콘텐트에 대한 사용자 관심이 시간에 따라 고정된 값을 가지거나 혹은 단순 증가/감소하는 것이 아니라, 대규모의 관심이 단시간에 집중되며 시간에 따른 쇠퇴가 급속히 진행되는 수명주기 상의 차이가 존재한다. 따라서, 소셜 뉴스 집적 서비스에서 사용자 관심을 향상시키는 방법을 도출하고 이를 적용하기 위해서는 먼저 뉴스 콘텐트에 적합한 수명주기 모델이 제시될 필요가 있다.

소셜 뉴스 집적 서비스에서의 뉴스 수명주기 모델이 가지는 중요성에 비해, 소셜 뉴스 집적 서비스와 관련된 현재까지의 연구는 대부분 사용자 관심의 성장 과정 및 그 원인에 집중하고 있다[13]. 즉, 기존의 관련 연구는 소셜 뉴스 집적 서비스에서 뉴스 콘텐트가 가지는 전체 수명주기를 대상으로 한 것이 아니라, 특정 뉴스 콘텐트가 사용자들의 소셜 네트워크를 통해 대규모 관심을 획득하게 되는 초기 과정만을 분석 대상으로 한다.

따라서, 본 연구는 뉴스 수명주기 패턴을 분석하여 급격하게 변화하는 사용자의 관심을 정확하게 예측할 수 있는 통계적 모델 제시를 목적으로 한다. 본 연구에서 제시하는 수명주기 모델은 뉴스 콘텐트가 가지는 전체 수명주기를 대상으로 한 것이며, 사용자의 관심이 단시간 내에 급증 혹은 급감하는 뉴스 콘텐트에 적합하다. 본 연구에서 제시된 모델을 기반으로 사용자 관심 증가를 위해 어떤 뉴스를 어떤 순서에 의해 얼마 동안 노출할 것인지에 대한 정확한 알고리즘 산출이 가능하다. 이는 뉴스 콘텐트의 유통을 활성화시킬 뿐 아니라, 다양한 사용자들 사이의 협력적 평가를 가능하게 하여 뉴스 콘텐트 평가의

정확성을 증가시키는 역할을 수행한다. 즉, 본 연구에서 제시하는 수명 시간 모델은 독자의 관심을 지속시키면서 다양한 콘텐트를 공급하려는 소셜 뉴스 집적 서비스에 매우 유용하게 적용될 수 있다.

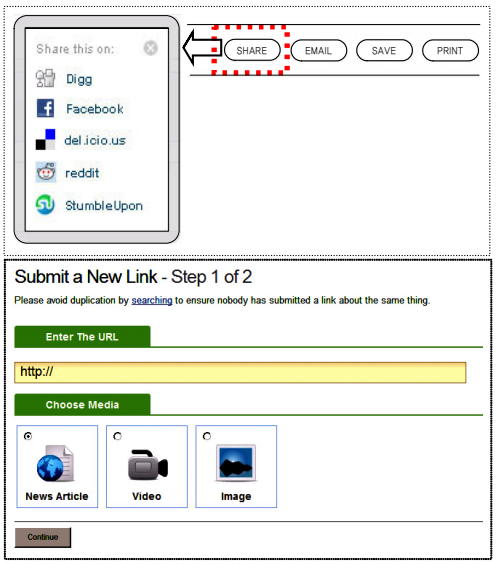
본 논문의 구성은 다음과 같다. 제 2장에서는 연구의 대상이 되는 Digg.com의 소셜 뉴스 집적 서비스에 대해 알아보며, 제 3장은 기존의 관련 연구를 살펴보도록 한다. 제 4장에서는 본 연구에서 도출된 모델 및 실험 결과를 설명하였으며, 마지막으로 제 5장에서 본 연구의 결론 및 추후 연구 방향을 제시하였다.

2. Digg.com의 소셜 뉴스 집적 서비스

본 연구는 소셜 뉴스 서비스 제공자 중 현재 가장 높은 인지도를 가지고 있으며[1], 기존 소셜 뉴스 집적 서비스 연구에서 주로 분석되고 있는 Digg.com을 대상으로 한다[14, 20]. 먼저 본 연구의 대상이 되는 Digg.com의 소셜 뉴스 집적 서비스에 대해 살펴보면 다음과 같다.

2.1 서비스 구성

Digg.com는 사용자들에 의해 뉴스가 선별되는 대표적인 소셜 뉴스 집적 서비스이다. Digg.com에 제출(Submit)된 뉴스 콘텐트를 스토리(Story)라고 하는데, 뉴스 제공자가 생성한 뉴스를 Digg.com에서의 인터페이스 혹은 뉴스 제공자의 인터페이스를 통하여 Digg.com

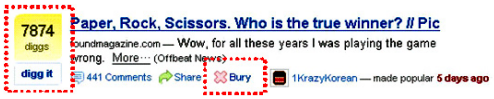


〈그림 1〉 Digg.com의 뉴스 제출 인터페이스 및 CNN에서의 뉴스 공유(Share) 인터페이스

으로 제출이 가능하다.

<그림 1>의 위 부분은 Digg.com 자체가 제공하는 뉴스 제출 사용자 인터페이스 화면이며, <그림 1>의 아래 부분은 CNN 사이트에서 직접 Digg.com 및 Reddit.com 등 다른 소셜 서비스들로 뉴스를 제출하는 공유 인터페이스를 나타내고 있다.

Digg.com에 제출된 스토리의 좋고 나쁨을 일반 사용자들은 <그림 2>와 같은 인터페이스를 통해 평가하게 된다. Digg.com에서는 뉴스의 찬성을 'digg'이라고 하며 스토리의 왼쪽에 있는 'digg' 버튼을 통해 투표할 수 있다. 제출된 스토리에 대한 반대 의사는 스토리 아래쪽에 있는 'Bury' 버튼을 통해 제출될 수 있다. 즉, 사용자는 이러한 digg/bury를 통해 스토리에 대한 평가를 실시한다. Digg.com에 제출된 스토리는 처음에 제출자(Submitter) 1인에 대한 digg값 1을 가지며, <그림 2>의



〈그림 2〉 Digg.com의 평가 인터페이스

스토리는 현재 7874건의 digg을 얻었음을 알 수 있다.

특정 스토리에 대한 이러한 'digg 수' 증가는 곧 해당 스토리에 찬성하는 사용자 증가를 의미하므로, 사용자 관심을 측정하는 지표로 사용될 수 있다. 즉, Digg.com의 사용자는 서비스의 콘텐트를 결정할 뿐 아니라, 집적 서비스에 '배달' 되어온 뉴스를 평가 하는 기능을 수행하며, 뉴스 콘텐트가 사용자들로부터 받는 관심의 정도는 해당 스토리의 'digg 수'를 통해 측정될 수 있다.

Digg.com에 제출된 스토리는 처음에는 공개 상태(Upcoming)로 일정 기간 머물러 있다가, 특정 조건을 만족할 경우 인기 상태(Popular)로 변경된다. 공개 상태의 스토리가 이러한 인기 상태로 변경되는 것을 스토리의 진급(Promotion)이 발생하였다고 하며, 스토리의 진급 이후 해당 스토리에 대한 'digg 수'가 급속히 증가한다[18]. 이는 Digg.com 사이트에서 공개/인기 상태의 스토리가 서로 다른 탭(Tab)으로 구분되어 있으며, 시작화면에서는 인기 상태의 스토리만을 보여지기 때문에, 인기 상태의 스토리에 사용자의 관심이 집중되기 때문이다.

<그림 3>에서 인기 상태와 공개 상태로 스토리를 구분하여 보여주는 Digg.com의 시작화면을 확인할 수 있다. 단, 공개 상태의 스토리를 인기 상태로 변경하는 이러한 Digg.com의 규칙 혹은 정책은 현재 공개되어 있지 않고 있으며, 일반적으로 특정 digg수 이상을

〈그림 3〉 Digg.com의 인기(Popular) 및 공개(Upcoming) 스토리 구분

얻게 되면 변경된다고 알려져 있다[18].

2.2 뉴스 카테고리

2007년 9월 기준으로 Digg.com은 뉴스 콘텐트를 7개의 대분류와 45개의 소분류로 구분하여 서비스하고 있으며, 이러한 대분류와 소분류를 각각 컨테이너(Container) 및 토픽(Topic)이라 한다.

컨테이너는 과학기술(Technology), 세계와 비즈니스(World and business), 비디오(Videos), 과학(Science), 연예(Entertainment), 게임(Gaming), 스포츠(Sports) 등의 총 7가지로 구성되어 있으며, 토픽은 모두 45가지 항목으로 이루어져 있다.

Digg.com의 2007년 9월의 카테고리 및 토픽 정보를 정리하면 <표 1>과 같다. <표 1>의 '세계와 비즈니스' 컨테이너에 있는 '2008 미대선(2008 U.S. Elections)' 토픽에서 알 수 있듯이, Digg.com의 컨테이너와 토픽은 특정 시기를 반영하며 시간에 따라 조금씩의 변화가 발생하기도 한다.

3. 관련 연구 분석

소셜 뉴스 집적 서비스에서 사용자의 소셜 네트워크(Social network)를 통해 사용자 관심이 확산되는 과정이 연구되었다. Lerman [13]은 소셜 뉴스 집적 서비스에서 사용자들

〈표 1〉 Digg.com의 카테고리 정보(2007년 9월)

컨테이너(Container)	토픽(Topic)
과학기술(Technology)	Apple, Design, Gadgets, Mods, Hardware, Tech News, Security, Linux/Unix, Microsoft, Software, Programming, Tech Deals
세계와 비즈니스 (World and Business)	Business and Finance, Politics, 2008 U.S. Elections, Political Opinion, World News, Offbeat News
비디오(Videos)	Animation, Comedy, Sports, Educational, Gaming, Music, People
과학(Science)	Space, Environment, Health, General Sciences
연예(Entertainment)	Celebrity, Movies, Music, Television
게임(Gaming)	Gaming News, Playable Web Games
스포츠(Sports)	Baseball, Basketball, Extreme, Football, Golf, Hockey, Motorsport, Soccer, Tennis, Other Sports

사이의 추천을 통해 특정 컨텐트에 대한 사용자들의 관심이 확산된다는 모델을 제시하였다. Lerman은 Digg.com에 제출된 뉴스 콘텐트가 사용자들로부터 받는 관심의 정도를 해당 스토리의 'digg 수'로 측정하였으며, 새로 전송된 뉴스 콘텐트가 시간에 따라 공개 상태 및 인기 상태로 각각 변화하는 과정을 사용자의 소셜 네트워크 크기 및 '시간당 digg 수'로 모델링 하였다. 여기서 특정 사용자의 소셜 네트워크는 해당 사용자의 'digg한 뉴스 목록'을 모니터링 하는 사용자 집단을 말하며, 소셜 네트워크가 큰 사용자는 자신을 모니터링 하는 사용자가 많기 때문에 그만큼 더 큰 영향력을 지니는 사용자가 된다. 따라서, 영향력이 큰 사용자가 제출하거나 혹은 영향력이 큰 사용자에 의해 digg된 뉴스 콘텐트는 그렇지 않은 콘텐트에 비해 공개 상태 및 인기 상태로 변화할 가능성이 높다고 분석되었다.

Lerman은 또한 앞서 제시한 사용자의 소셜 네트워크 모델을 활용하여 특정 뉴스 콘텐트에 대한 품질을 예측하는 알고리즘을 제시하였다. 즉, 뉴스가 공개 상태 혹은 인기 상태로 진급할지 여부를 해당 콘텐트를 Digg.com에 제출한 사용자의 네트워크 크기 및 초기 '시간당 digg 수' 등의 정보에 의해 예측하는 알고리즘을 제시하였다.

사용자들 사이의 관계를 기반으로 한 이러한 연구는 소셜 뉴스 집적 서비스에서 초기 뉴스 확산 과정에 대한 인과 관계를 설명하는 반면, 성장된 이후의 뉴스 스토리에 대한 수명주기 가 연구 범위에 포함되지 않는다는 한계를 가지고 있다.

뉴스 콘텐트 이외의 웹 문서, 블로그 등의 다른 웹 콘텐트에 대한 수명주기 연구가 활발히 진행 중에 있다[10, 17, 3]. 특별히 Chen[2]은 소셜 뉴스 집적 서비스가 아닌 일반적인 웹 콘텐트 제공 서비스를 대상으로 뉴스 콘텐트에 관한 사용자의 관심 변화 모델을 제시하였다. Chen은 특정 뉴스와 관련된 문서를 이벤트로 정의하고 이벤트 관련 문서의 수와 분포를 통해 수명주기를 모델링 하였다.

Chen의 이러한 모델은 시간 변화에 따른 뉴스에 대한 관심의 변화를 모델링 했다는 점에서 본 연구와 공통점을 지니나, 다른 웹 콘텐트에 대한 연구와 마찬가지로 사용자 관

심의 급격한 변화가 발생하는 소셜 뉴스 집적 서비스에서는 적합하지 않는다는 한계점을 지니고 있다.

4. 뉴스 수명주기 패턴 분석

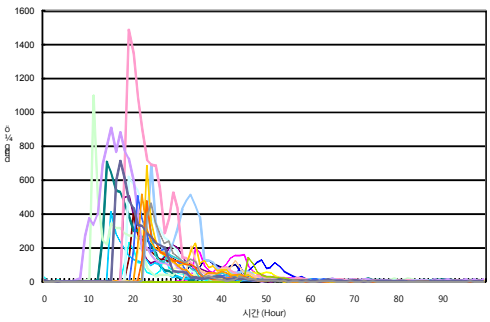
본 연구는 2007년 5월부터 9월까지 Digg.com의 인기 스토리 13,599개를 대상으로 다음과 같은 분석을 실시하였다. 하나의 스토리 별로 1시간 단위로 100시간(5일)동안 이벤트 정보('digg 수' 및 '의견 수')를 수집하였다. <표 2>는 수집된 스토리들의 카테고리 별 분포를 정리한 것으로, Digg.com은 과학기술, 세계와 비즈니스 및 비디오 카테고리의 스토리 비중이 상대적으로 높음을 알 수 있다.

Digg.com의 뉴스 수명주기 패턴을 분석하기 위해, 먼저 특정 토픽에 속한 스토리들의 'digg 수' 변화를 시간에 따라 나타내면 다음의 <그림 4>와 같다.

<그림 4>는 과학기술 컨테이너의 소프트웨어 토픽에 속한 스토리들에 대한 수명주기를 나타낸 것으로, 스토리가 제출된 시점을 '0'이라고 했을 때, 제출 시점부터 100시간까

〈표 2〉 카테고리별 스토리 분포

카테고리	개수	비율(%)
과학기술	4,010	29.5
세계와 비즈니스	3,776	27.8
비디오	2,280	16.8
과학	1,568	11.5
연예	876	6.4
게임	726	5.3
스포츠	363	2.7
전체 카테고리	13,599	100.0



〈그림 4〉 과학기술 컨테이너의 소프트웨어 토픽에 포함된 뉴스들의 수명주기

지의 '시간 별 digg 수'를 나타내고 있다. <그림 4>에서 알 수 있듯이, 패턴이 불규칙하여 일반적인 특성을 찾기 어렵지만, 대체적으로 '시간별 digg 수'가 급격히 증가하는 구간이 존재하며, '시간별 digg 수'가 최고점에 이르러서는 급격히 감소하는 패턴이 존재한다는 것을 알 수 있다. 본 연구에서는 '시간별 digg 수'로 클러스터링을 실시하여 이러한 패턴들의 공통점에 대해 좀 더 자세히 살펴보기로 한다.

4.1 Kohonen SOM을 이용한 클러스터링

뉴스의 수명주기 패턴을 파악하기 위하여 먼저 Kohonen SOM(Self-Organizing Map)을 이용한 클러스터링을 실시하였다[12]. SOM은 신경망 분야의 하나로 입력 벡터들의 위상적 순서(Topological order)를 유지하도록 학습이 진행된다는 특징을 가지고 있으며, 위상 순서 보존이 필요한 패턴 분석 연구에 적용하여 그 성능을 검증 받았다[4].

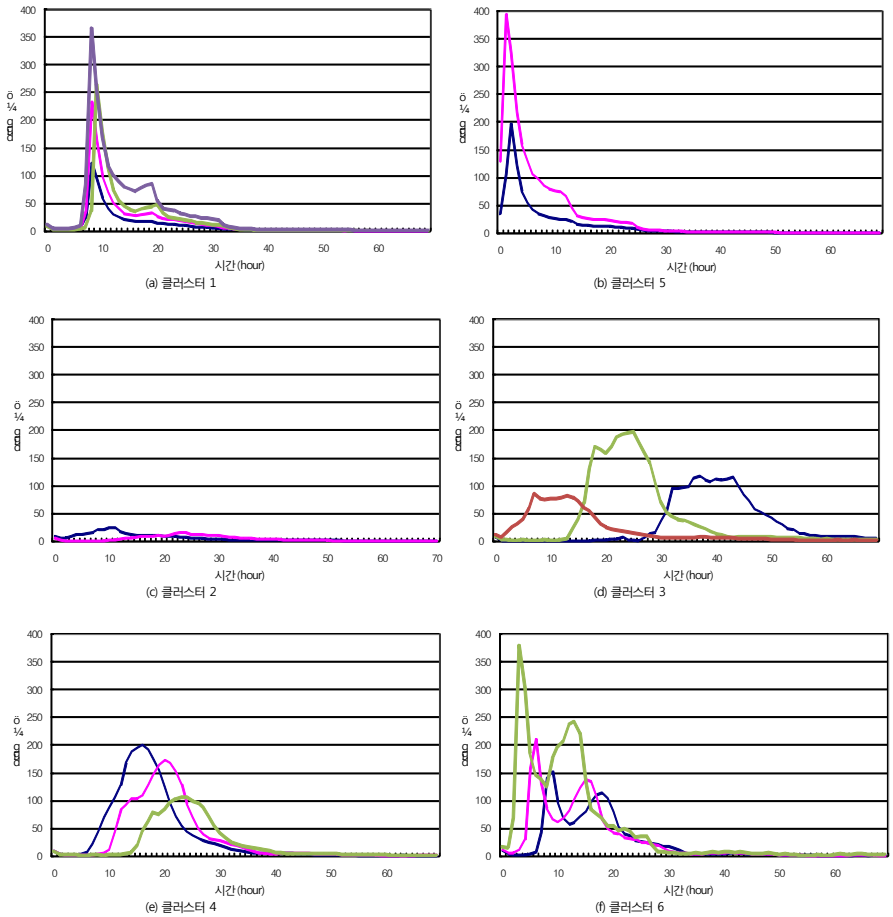
'시간 별 digg수'를 입력 값으로 하여, 9 x 9 그리드 환경에서 클러스터링 하였으며, 산출된 이웃 클러스터(Nearest cluster)을 수명주

〈표 3〉 뉴스 수명주기 패턴 별 클러스터링 결과

ID	설 명	(%)
1	한번 최대 값을 가졌다가 급격히 감소	76.7
2	전체적으로 평평한 구조	13.6
3	뚜렷한 변곡점을 보이지 않고, 증가/감소의 형태를 띠는 구조	3.6
4	비교적 작은 최대 값을 가지며, 봉우리 형태를 띠는 구조	2.1
5	생성직후 약 2시간 이내로 최대 값을 가졌다가 급격히 감소	2.1
6	두 번의 변곡점을 가지는 구조	1.9

기 형태에 따라 재 분류하여 <표 3>과 같은 6개의 클러스터를 도출하였다.

<그림 5>는 각 클러스터에서 임의로 선택된 몇 가지 뉴스 샘플들에 대한 '시간 별 digg 수'를 표시하고 있다. 클러스터 1과 클러스터5가 전체 스토리에서 차지하는 비율은 약 78.7%로 가장 큰 클러스터에 해당한다. <표 3>과 <그림 5>에서 알 수 있듯이, 이러한 클러스터 1과 클러스터 5는 스토리 생성 후 한번 최대 값을 가진 다음 급격히 감소하는 패턴을 가지는 특성이 있다. 두 번째로 많



〈그림 5〉 뉴스 수명주기 패턴 별 클러스터링 결과

은 비율을 차지하는 클러스터 2는 전체 스토리의 약 13.6%를 차지하는데, 최대 digg수 자체가 50미만으로 상당히 작다는 특성을 지닌다. 즉, 클러스터 2는 최대 값으로 급격한 성장의 부분이 없다는 점에서 클러스터 1과 차이가 있으나, 쇠퇴된 이후 부분은 유사하다고 볼 수 있다. 클러스터 3과 4는 '시간당 digg 수'의 최대 값은 클러스터 1과 비교해 상대적으로 작으나, 급격한 증가/감소의 패턴을 지니고 있다는 점에서는 동일하다. 마지막으로 클러스터 6은 급격한 성장 및 쇠퇴 이후에 다시 성장하는 패턴을 지니고 있다는 점에서 다른 클러스터들과 차이가 있으나, 이후 급격한 쇠퇴로 수명 시간이 끝난다는 점에서는 동일하다.

따라서, 앞서의 패턴 별 클러스터링의 결과에서 Digg.com의 뉴스 콘텐트는 최대값의 차이는 있으나 대부분 수명주기가 '시간 별 digg 수'가 최대값을 가질 때까지 급격히 증가하는 기간과 '시간 별 digg 수'가 급격히 하강하는 기간이 존재함을 알 수 있다. 즉, 위상 순서 유지 속성을 지니는 SOM을 통해 '시간 별 digg 수'의 유사 패턴 별로 클러스터링된 클러스터의 분석결과 '시간별 digg 수'의 급격한 증가 및 감소가 대부분의 클러스터에서 공통적으로 발견될 수 있었다.

4.2 뉴스의 성장 특성 분석

사용자의 관심이 최대로 집중되는 시점을 '시간당 digg 수'가 최대가 되는 시점으로 정의하였을 때, 이러한 시점은 각 스토리 별로 차이가 존재한다.

이러한 사용자 관심의 급격한 성장 원인을 소셜 뉴스 집적 서비스의 콘텐트 게시 정책과 같은 외부적 요인과 콘텐트 품질, 소재의 참신성 등과 같은 스토리 자체가 지니는 내부적 속성으로 구분해 볼 수 있다.

〈표 4〉 인기 스토리가 된 시점 분석

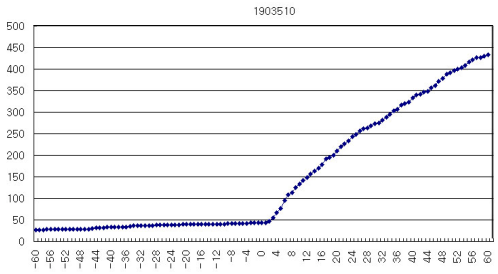
설 명	시간
인기 스토리가 되기까지 걸린 시간의 평균	14.22
'시간당 digg 수'가 최대가 되기까지 걸린 시간의 평균	15.15
두 시간 차이의 평균	0.93
RMSE	1.72

4.2.1 뉴스 성장의 외부적 요인

먼저 외부적 요인에 대해 살펴보면 다음과 같다. 스토리의 제출 시점부터 인기 스토리가 되기까지 걸린 시간, '시간당 digg 수'가 최대가 되기까지 걸린 시간 및 두 시간 차이 각각에 대한 평균 등을 정리하면 <표 4>와 같다.

Digg.com에서 특정 스토리가 공개 상태에서 인기 상태로 진급된 이후 평균 0.93시간 이내의 시점에 최대 '시간당 digg 수' 값을 가지는 것으로 나타났다. 인기 스토리가 된 시각과 시간당 최대 digg수 값을 가지는 시각 사이에 비교적 큰 관련성이 존재하며, 이는 스토리의 진급이 사용자 관심의 급격한 성장을 발생시킨다는 것을 의미한다. <그림 6>은 스토리 id '1903510'의 인기 스토리(popular story)가 된 시점을 '0'으로 놓고, 전후 1시간의 '분단위 digg 수'를 그래프로 나타낸 것으로 이러한 스토리의 인기 상태 변화와 사용자 관심의 성장 사이의 관련성을 확인할 수 있다. 즉, 앞서 언

급한 바와 인기 스토리가 된 시점 0부터 '분단위 digg 수'가 급격히 상승하고 있음을



〈그림 6〉 스토리 id '1903510'의 'digg 수' 변화

알 수 있다.

따라서, 위의 <그림 6>에서 스토리의 진급이 사용자 관심의 급격한 성장을 발생시킨다는 것을 다시 확인할 수 있다.

다음으로 스토리 자체가 지니는 내부적 속성과 뉴스 성장과의 관계를 살펴보도록 하자.

4.2.2 뉴스 성장의 내부적 요인

사용자들이 협력하여 뉴스의 내용을 평가하고 이를 확산시키는 소셜 뉴스 집적 서비스에서, 소재가 참신하거나 우수한 콘텐트 품질을 가지는 스토리의 경우 일반적으로 많은 수의 digg 혹은 의견이 발생한다. 또한 앞서의 분석에서 인기 스토리가 되는 시점과 '시간당 digg 수'가 최대가 되는 시간 사이의 차이가 평균 0.93시간 미만이므로, 스토리 자체가 지니는 내부적 속성과 뉴스 성장이 밀접한 관계에 있다면 인기 스토리가 되는 시간과 'digg 수' 사이에 상관 관계가 있어야 한다.

<표 5>는 인기 스토리로 상태가 변경되는 시점까지의 누적 'digg 수'의 평균, 최소값 및 최대값 등을 정리한 것으로, 누적 'digg 수'가 최소값 8에서부터 최대값 935까지 그 범위가 다양할 뿐 아니라, 상대적으로 큰 RMSE값에서 digg수와 인기 스토리가 된 시점간에 상

〈표 5〉 인기 스토리가 된 시점까지 누적 'digg 수'

설명	개수
평균	51.5
최소값	8.0
최대값	935.0
RMSE	59.83

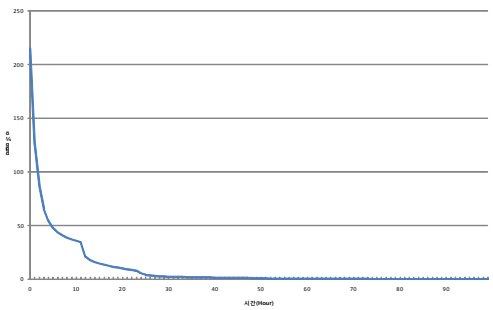
관관계가 약하다는 것을 알 수 있다. 각 뉴스에 달리는 의견 수도 위의 누적 'digg 수'와 마찬가지로 뚜렷한 상관관계를 찾을 수 없다. 즉, 어떤 스토리가 일정 'digg 수'를 얻어 인기 스토리로 등극하는 것에는 스토리 자체가 지니는 내부적 속성보다는 Digg.com의 내부적 판단과 같은 외부적 요인일 가능성이 크다고 판단된다.

Digg.com에 제출된 스토리가 공개 상태에서 인기 스토리로 그 상태가 변경되면, Digg.com의 첫 페이지(Front page)에 노출되어 사용자들의 집중적인 관심을 받기 시작한다. 따라서, 시간당 최대 'digg 수' 값을 갖는 시점은 결국 스토리의 속성이 아니라 Digg.com의 콘텐트 게시 정책에 의해 결정된다고 추측된다.

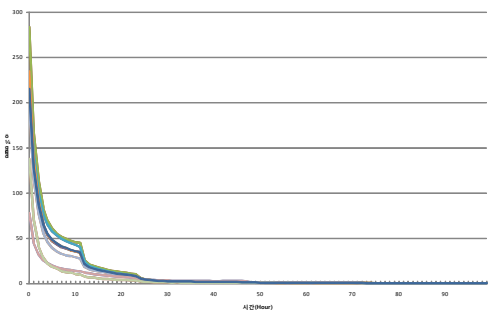
4.3 뉴스의 쇠퇴 패턴 분석

본 연구에서 사용된 모든 스토리를 대상으로 인기 스토리가 된 시점부터 100시간까지 시간당 평균 digg수를 나타내면 <그림 7>과 같다.

<그림 7>에서 나타난 바와 같이, 인기 스토리가 된 직후 1시간 동안의 평균 'digg 수'는 214.97으로 최대 '시간당 digg 수'를 가지게 된다. 최대 '시간당 digg 수'를 가진 이후의 1시간 동안 'digg 수'는 127.76으로 40.6%의 감소가 발생한다. 또한 이후 각 시간 별로



〈그림 7〉 (전체 카테고리)'시간당 digg 수'의 감소 패턴



〈그림 8〉 (각 카테고리별)'시간당 digg수'의 감소 패턴

차례로 19.1%, 9.7%, 5.1%의 감소가 발생하여, 4시간 동안의 총 감소량은 전체의 75%를 차지하며, 인기 스토리가 된지 12시간이 지났을 때는 최대 사용자 관심의 약 90%를 잃게 된다.

즉, 사용자 관심의 감소는 인기 스토리가 된 이후 1시간 이내에 가장 크게 발생하며, 4시간 뒤에는 약 75%의 누적 감소가 발생한다. '시간당 digg 수'가 1미만이 되는 시점은 인기 스토리가 된 후 평균 51시간이 되는 시점이다. 따라서, 인기 스토리로 상태가 변경되어 한 시간 이내에 최대 '시간당 digg 수'를 가지게 되는 뉴스 스토리는 이 후 1시간 이내에 사용자 관심의 급격한 감소가 시작되어 4시간 이내에 최대값의 약 25% 수준으로 감소하게 됨을 알 수 있다.

<그림 8>과 <표 6>을 통해 스포츠 및 비디오 카테고리와 다른 5개의 카테고리 사이의 차이가 다소 존재함을 알 수 있다. 즉, 최대 '시간당 digg 수'의 평균 값이 다른 5개의 카테고리와 비교하여 118개 이상 적으며, 'digg 수'가 보다 더 신속하게 감소되는 것을 그래프 상에서 확인할 수 있다. 반면 '시간당 digg 수'가 급격히 감소하는 패턴은 이들 두 개의 카테고리도 또한 동일함을 알 수 있다.

따라서, 본 연구에서는 이러한 뚜렷한 특성을 지니는 뉴스의 쇠퇴 패턴을 모델링하기 위해, '시간당 digg 수'를 로그함수로 평활화(Smoothing)한 후, 선형 회귀 분석을 적용하기로 한다. 이는 '시간당 digg 수'가 급격히 감소하는 형태를 보이고 있는 점에 착안한 것이며, 아래 수식과 같이 로그 함수를 활용하여 충분한 평활화를 실시한 후 선형 회귀 분석을 실시하였다.

$$\ln(\ln(d_{k,\ t})) = a_k t + b_k$$

위 식에서 k는 카테고리 id이며, t는 인기 스토리가 된 이후의 시간 값으로 0이상의 정

〈표 6〉 카테고리별 최대 '시간당 digg 수'의 평균

카테고리	최대 '시간당 digg 수'의 평균
과학기술	233.01
세계와 비즈니스	283.67
비디오	77.89
과학	218.07
연예	234.68
게임	196.76
스포츠	137.51
전체 카테고리	214.97

수이다. 또한, $d_{k,\ t}$는 시간 t에서의 카테고리 k의 '시간당 digg 수', a_k와 b_k는 각각 카테고리 k에서의 1차 선형회귀 모형의 모수이다. 평활화를 실시하는 다양한 기법이 가능하나[20], 본 연구에서는 다양한 테스트를 통해 상대적으로 성능이 우수한 위의 방식을 채택하였다.

이러한 평활화 전처리 과정 후 카테고리별 회귀 분석을 실시한 결과를 정리하면 <표 7>과 같다. 본 연구에서는 회귀 모델의 정확성 측정을 위해 일반적으로 사용되는 RMSE, 결정계수(R-Square) 및 수정 결정계수(Adjusted R-Square) 지표를 사용하였다. 수정 결정계수는 데이터의 개수가 많아지면 결정계수의 값이 1에 가까워지는 단점을 보완하는 지표이다[11].

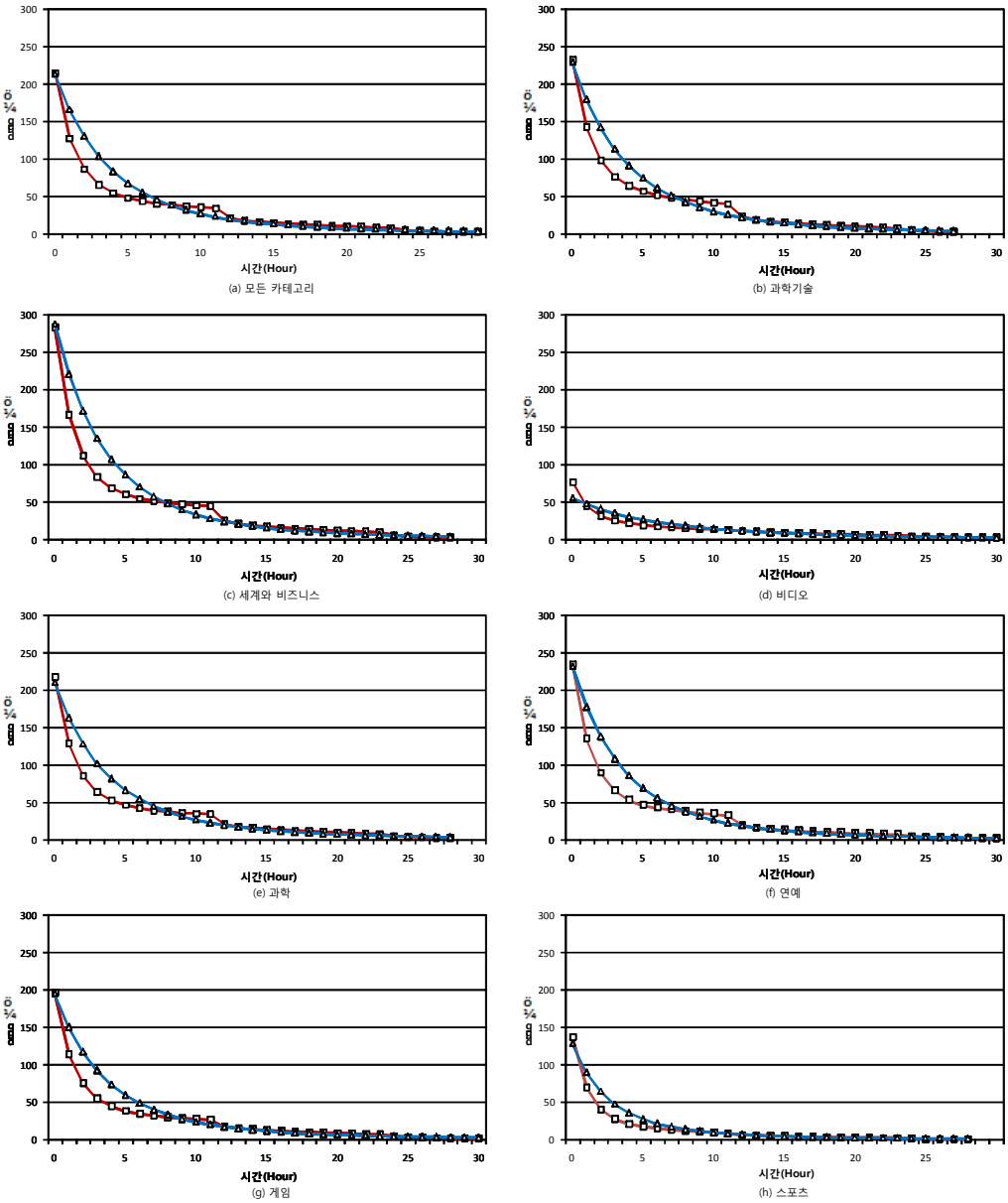
모든 카테고리 각각에 대해서 유의 확률이 0.0001미만이며, RMSE의 최대값 0.31을 실제 'digg 수'로 변환하면 3.9개이기 때문에, 상당히 정확한 회귀 모델이 산출되었음을 <표 7>에서 확인할 수 있다. 즉, 사용자들의 관심이 최대로 집중된 시점 이후의 쇠퇴 과정을 뉴스의 카테고리 별로 비교적 정확히 예측할 수 있는 모델이 산출되었다고 판단할 수 있다.

<표 7>에서 모수 a_k는 '시간당 digg 수' 변화 량을 의미하며, 모수 b_k는 최대 '시간당 digg 수'를 나타낸다. 앞서의 <그림 8>과 <표 6>에서와 마찬가지로 <표 7>에서도 동일하게 스포츠 및 비디오 카테고리가 다른 5개의 카테고리와 구별되는 특성이 존재함을 알 수 있다. 즉, 최대 '시간당 digg 수'를 나타내는 모수 b_k의 값이 비디오 카테고리의 경우 다른 5개의 카테고리와 비교해 최소 0.27이상 더 작은 값을 가지며, 스포츠 카테고리는 최소 0.09이상 더 작다. 뿐만 아니라, 스포츠 카테고리의 경우 '시간당 digg 수' 변화 량을 나타내는 모수 a_k의 절대 값이 다른 5개의 카테고리에 비해 최소 0.02이상 더 큰 값을 가진다. 즉, 스포츠 및 비디오 카테고리는 최대 '시간당 digg 수'가 다른 카테고리에 비해 작으며, 특별히 비디오 카테고리의 경우 다른 카테고리에 비해 상대적으로 '시간당 digg 수'의 감소가 크다고 판단된다.

마지막으로, <그림 9>는 앞서 제안된 회귀 모델에 의해 산출된 예측 값(범례 △)과 실제

〈표 7〉 카테고리별 선형 회귀 분석 결과

카테고리	b_k	a_k	결정계수 (R-Square)	수정 결정계수 (Adjusted R-Square)	RMSE	유의 확률
과학기술	1.6929	-0.0460	0.9435	0.9413	0.0910	< 0.0001
세계와 비즈니스	1.7339	-0.0476	0.9148	0.9117	0.1215	< 0.0001
비디오	1.3920	-0.0390	0.7688	0.7640	0.3086	< 0.0001
과학	1.6772	-0.0485	0.9261	0.9234	0.1147	< 0.0001
연예	1.6946	-0.0499	0.9520	0.9504	0.1035	< 0.0001
게임	1.6628	-0.0509	0.9220	0.9194	0.1323	< 0.0001
스포츠	1.5807	-0.0762	0.9053	0.9018	0.2063	< 0.0001
전체 카테고리	1.6800	-0.0483	0.9386	0.9364	0.1070	< 0.0001



〈그림 9〉 인기 스토리가 된 시점 이후의 시간당 평균 'digg 수'

데이터 값(범례 □)을 각각의 카테고리 별로 비교한 결과를 보여준다. 앞서 설명한 바와 같이 인기 스토리로 상태가 변경되어 최대 '시간당 digg 수'를 가지게 되는 뉴스 스토리는 이 후 4시간 이내에 최대 값의 약 25% 수준까지 '시간당 digg 수'의 급격한 감소가 발생하여 인기 스토리가 된지 12시간이 지났을 때는 90%의 사용자 관심을 잃게 된다. 또한

뉴스의 카테고리 별로 '최대 시간당 digg 수'와 '시간당 digg 수'의 감소 패턴 상의 차이가 존재한다. 본 연구에서 제안한 모델이 이러한 사용자 관심의 변화 및 카테고리 별 감소 패턴의 차이를 정확히 모델링 한다는 것을 <그림 9>에서 확인할 수 있다.

5. 결 론

본 논문은 인터넷 뉴스의 수명주기를 분석하여 성장 및 쇠퇴 과정에 대한 모델을 제시하였다.

먼저, 특정 뉴스 스토리가 급속히 성장하는 데에는 콘텐트의 내부 속성보다는 콘텐트 게시 정책과 같은 외부적인 요인이 크게 작용하며, 공개 상태에서 인기 상태로 진급된 이후 평균 0.93시간 이내의 시점에 최대 '시간당 digg 수' 값을 가지는 것으로 분석되었다. 즉, 특정 뉴스 스토리가 사용자 소셜 네트워크에 의해 인기 상태로 진급한 뒤[13], 비교적 짧은 시간 내에 최대한의 사용자들의 관심을 가지게 됨을 알 수 있다.

또한 본 연구에서는 뉴스의 쇠퇴 패턴을 분석하여 사용자 관심의 변화를 예측할 수 있는 통계적 모델이 제시되었다. 이는 기존의 사용자 소셜 네트워크 기반의 연구를 보완하는 것으로, 사용자의 관심이 시간에 따라 급속히 쇠퇴하는 뉴스 콘텐트의 수명주기 패턴을 정확하게 예측 가능하게 한다. 특별히 각각의 뉴스 카테고리 별로 산출된 모델에 의해, 비디오 카테고리의 경우 최대 '시간당 digg 수'의 크기가 다른 스토리의 약 1/3 수준으로 사용자들의 관심을 비교적 덜 받고 있으며, 관심의 감소 또한 급격함을 분석할 수 있었다.

뉴스 카테고리 별 사용자 관심의 변화 량이 본 연구에서 제안된 수명주기 모델에 의해 예측 가능하므로, 이를 통해 소셜 뉴스 집적 서비스의 시작 화면에 어떤 뉴스를 얼마 동안 게시할지를 사용자 관심의 극대화 측면에서 판단할 수 있다[9, 21]. 즉, 사용자 관심의 극대화를 위해 웹 페이지의 링크 순서 및 콘텐트의 나열 순서를 정렬하는 기존의 방법에 본 연구에서 산출된 콘텐트 수명주기 모델을 적용할 수 있으며, 이러한 적용을 통해 대규모의 관심이 단시간에 집중되며 시간에 따른 쇠퇴가 급속히 진행되는 뉴스 콘텐트에 적합한 정렬 결과를 도출할 수 있다. 이는 사이트 방문수 및 광고 수입의 향상뿐 아니라 사용자들의 뉴스 콘텐트 평가 참여의 증가를 통한 서비스 품질의 향상까지 기여할 수 있다.

본 연구에서 제시하는 수명 시간 모델은 독자의 관심을 지속시키면서 다양한 콘텐트를 공급하려는 소셜 뉴스 집적 서비스에 매우 유용하게 적용될 수 있다. 아울러, 본 연구에서 사용된 분석 방법 및 도출된 모델 등을 활용하여, 블로그 포스트 및 동영상 등 다른 웹 리소스의 수명주기에 대한 추후 연구가 가능할 것으로 예상된다.

참 고 문 헌

[1] Alexa.com, http://www.alexa.com.

[2] Chen, C. C., Chen, Y. T., and Chen, M.

C., “An Aging Theory for Event Life Cycle Modeling,” *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS-PART A : SYSTEMS AND HUMANS*, Vol. 7, No. 2, 2007, pp. 237-248.

[3] Chen, C. C. and Chen, M. C., PVA : A Self-Adaptive Personal View Agent, *Journal of Intelligent Information Systems*, Vol. 18, No. 2, 2002, pp. 173-194.

[4] Chicco, G., Napoli, R., and Pigilone, F., “Load pattern clustering for short-term load forecasting of anomalous days,” *IEEE Porto Power Tech Conference*, Porto, Portugal, 2001.

[5] CNN, http://www.cnn.com.

[6] Digg.com, http://www.Digg.com.

[7] Digg.com API, http://apidoc.Digg.com.

[8] Flavia'n, C. and Gurrea, R., “Reading newspapers on the Internet : the influence of web sites' attributes,” *Internet Research*, Vol. 18, No. 1, 2008, pp. 26-45.

[9] Garofalakis, J., Kappos, P., and Mourloukos, D., “Web site optimization using page popularity,” *IEEE Internet Computing*, Vol. 3, No. 4, 1999, pp. 22-29.

[10] Gurzick, D. and Lutters, W. G., “From the personal to the profound : understanding the blog life cycle,” *Conference on Human Factors in Computing Systems*, Montréal, Québec, Canada, 2006, pp. 827-832.

[11] Hayter, A. J., *Probability and Statistics for Engineers and Scientists*, Duxbury Press, 2001.

[12] Kohonen., T., Hynninen, J., Hynninen, J., and Kangas, J., *SOM-PAK, The Self-Organizing Map Program Package, User's Guide*, Helsinki University of Technology, 1995.

[13] Lerman, K. and Galstyan, A., “Analysis of Social Voting Patterns on Digg,” *Proceedings of the first workshop on Online social networks*, Seattle, Washington, USA, 2008, pp. 7-12.

[14] Lerman, K., “Social Information Processing in Social News Aggregation,” *IEEE Internet Computing*, Vol. 11, No. 6, 2007, pp. 16-28.

[15] The New York Times, http://www.times.com.

[16] O'Reilly, T., “What is Web 2.0 : Design Patterns and Business Models for the next generation of software,” http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, 2005.

[17] Perseus, The Blogging Iceberg, Perseus Survey Solutions (http://www.perseus.com/blogsurvey), 2004.

[18] Saleem, M., “The Decline and Fall of Quality on Digg, http://www.readriteweb.com/archives/the_decline_and_fall_of_quality_on_digg.php, 2008.

[19] Reddit.com, http://www.Reddit.com.

[20] Wu, F. and Huberman, B. A., “Popularity, Novelty and Attention,” *Conference on Electronic Commerce*, Chicago, Il, USA, 2008, pp. 240-245.

[21] Wu, F. and Huberman, B. A., “The economics of attention : maximizing user value in information-rich environments, *International Conference on Knowledge Discovery and Data Mining*, San Jose, California, USA, 2007, pp. 16-20.

저 자 소 개

원미경 (E-mail : ovverplay@gmail.com)
2005년 아주대학교 미디어학부 (학사)
2008년 서울대학교 산업공학과 (석사)
2008년~현재 다음커뮤니케이션
관심분야 웹 시각화, 유저 인터페이스

이상진 (E-mail : sjinlee@snu.ac.kr)
1998년 서울대학교 산업공학과 (학사)
2000년 서울대학교 산업공학과 (석사)
2000년~2005년 핸디소프트, 수석연구위원
2005년~2006년 삼성네트웍스 기업솔루션사업팀, 과장
2007년~현재 서울대학교 산업공학과 박사과정 재학 중
관심분야 웹 검색, 추천 시스템

이승준 (E-mail : zaregn81@snu.ac.kr)
2007년 서울대학교 산업공학과 (학사)
2007년~현재 서울대학교 산업공학과 석사과정 재학 중
관심분야 웹 검색, 추천 시스템

박종헌 (E-mail : jonghun@snu.ac.kr)
1990년 서울대학교 산업공학과 (학사)
1992년 서울대학교 산업공학과 (석사)
1998년~2000년 Ph. D., Industrial and Systems Eng., Georgia Institute of Technology
2001년~2003년 Assistant Professor, Pennsylvania State University
2003년~2004년 KAIST 교수
2004년~현재 서울대학교 산업공학과 교수
관심분야 웹 공학