# UCC 비디오 서비스에서 소셜 네트워크를 통한 사용자 신뢰도 도출

## Evaluating the User Reputation through Social Network on UCC Video Services

조현철, Hyunchul Cho*, 한요섭, Yo-Sub Han**, 김래현, Laehyun Kim***

**요약** 최근 들어 사용자들이 직접 저작하고 이를 공유하는 UCC(User Created Content)가 급격히 증가하고 있다. 이에 따라 방대한 UCC를 사용자들에게 효과적으로 제공하기 위하여, 질이 낮은 UCC를 필터링하는 알고리즘이나 UCC의 검색 또는 추천 알고리즘에 대한 연구가 많이 진행되고 있다. 본 논문에서는 사용자에게 UCC 컨텐츠를 제공할 때 컨텐츠의 품질을 추정할 수 있는 요소로 사용자 신뢰도를 제안한다. 이를 위해 먼저 UCC 컨텐츠 제공 서비스 상에서 사용자 간의 소셜활동을 기반으로 소셜 네트워크를 구축하고, 사용자 신뢰도를 계산하기 위한 소셜 활동 정보를 추출한다. 그리고 소셜 네트워크를 통해 사용자 신뢰도를 계산하며, 다양한 소셜 정보 요소를 적용할 수 있는 확장 가능한 알고리즘을 제안한다.

**Abstract** Recently user-generated contents have been drastically increased. In this paper, we introduce the user reputation which can be used to evaluate quality of the content they created. First we have composed a social network that is based on user activity. And we have developed the algorithm to evaluate the users' reputation using this social network.

*해심어: User Reputation, Social Network, UCC*

본 연구는 지식경제부 및 정보통신연구진흥원의 IT핵심기술개발사업의 일환으로 수행하였음. [2008-S-024-01, RichUCC 기술개발]

*주저자 : 한국과학기술연구원 지능시스템연구본부 지능인터랙션연구센터 연구원 e-mail: hccho@kist.re.kr
**공동저자 : 한국과학기술연구원 지능시스템연구본부 지능인터랙션연구센터 선임연구원 e-mail: emmous@kist.re.kr
***교신저자 : 한국과학기술연구원 지능시스템연구본부 지능인터랙션연구센터 선임연구원 e-mail: laehyunk@kist.re.kr

## 1. 서론

최근 몇 년간 웹 콘텐츠의 제공 및 사용 형태가 빠르게 변화하고 있다. 초창기 웹 콘텐츠의 경우 기존 출판 미디어 방식대로 소수의 전문 공급자가 대부분의 콘텐츠를 만들어 보급하였고 다수의 일반 사용자는 제공되는 콘텐츠를 소비하는 방식이었다. 하지만 근래 들어 사용자 참여와 협동이 강조되는 웹 2.0의 도래로 말미암아, 일반 사용자들 역시 많은 콘텐츠 생성에 참여하게 되었고, 최근 몇 년간 블로그나 공유 사이트 등을 중심으로 제작자와 사용자의 명확한 구분이 없는 형태의 콘텐츠 저작/사용이 가속화되고 있다. 특히 예전에 텍스트 콘텐츠가 웹의 대부분을 차지했던 것과는 달리 이제는 Youtube나 Flicker 등의 비디오와 이미지 UCC 공유사이트가 전체 웹에서 차지하는 비중도 급격히 높아지고 있다.

UCC는 기존의 방식인 공급자 생성 컨텐츠와는 다르게 생성된 컨텐츠의 품질이 아주 다양한 모습을 보인다. 제한된 수의 공급자가 컨텐츠를 생성할 때는 각 컨텐츠간의 질적 차이가 비교적 크지 않았지만, UCC 컨텐츠들은 전문가 수준부터 아주 낮은 수준까지 분포가 다양하고 심지어는 저속한 내용의 컨텐츠들도 다수 포함된다. 이 때문에 사용자 생성 컨텐츠를 제공할 때에는 이러한 컨텐츠들의 필터링과 랭킹이 매우 중요하고 또한 복잡한 문제이다.

또한 UCC 비디오나 사진의 경우 텍스트 정보가 없거나 너무 적어서, 기존의 정보 검색에서 효과적으로 사용하던 키워드 기반의 알고리즘을 적용하기 어려운 문제가 있다. 기존의 영화 정보 제공 서비스나 비디오 컨텐츠 제공 서비스에서는 영화 제목이나 줄거리, 배우 정보, 수상 내역 등 풍부한 텍스트 정보를 공급자가 미리 입력하고 이를 통해 사용자가 쉽게 검색을 할 수 있도록 하거나, 장르나 시대 등의 기준으로 공급자가 미리 분류를 하여 사용자가 검색하는 것을 도왔다. 하지만 UCC 동영상은 사용자들이 직접 생성하므로, 기존 공급자 중심의 비디오 컨텐츠에서 제공하는 정형화된 메타데이터와 잘 정의된 분류를 통한 컨텐츠 제공을 기대하기 어려운 문제가 있다. 또한 하루에도 몇 십만 건씩 생성되는 UCC 동영상의 증가량으로 인해 인력으로 양질의 동영상을 찾아 제공하는 데에는 한계가 있다.

하지만 이러한 UCC 환경에서 사용자들은 컨텐츠를 생성할 뿐만 아니라, 사용자 간의 다양한 상호작용 또는 컨텐츠에 대한 상호작용 등이 동시에 일어나며 이러한 활동이 컨텐츠의 품질에 대한 단서를 제공하기도 한다. 그러므로 정보 검색 면에서는 기존에 사용하던 컨텐츠와 컨텐츠간의 링크에 더하여 다양한 사용자-컨텐츠 관계, 사용자-사용자 관계 등의 정보를 이용할 수 있게 되었다. 그러므로 특히 이러한 비텍스트 컨텐츠에 대한 검색 성능 향상을 위해 소셜 네트워크 기반 정보를 이용하여 검색 알고리즘을 보완하는 연구가 많이 진행되고 있다.

본 논문에서는 UCC 비디오 서비스에서 검색이나 추천을 통해 사용자에게 컨텐츠를 제공할 때 적합한 컨텐츠를 판별하기 위한 한 요소로써 사용자 신뢰도를 제안한다. UCC 비디오 서비스 상에서 사용자 간의 상호작용이나 컨텐츠에 대한 다양한 소셜활동을 통해 소셜 네트워크를 구성하고, 이를 통해 다른 사용자들로부터 많은 지지를 받는 신뢰도가 높은 사용자를 찾아내었다. 이런 사용자들은 기존에 좋은 컨텐츠를 많이 생성하였고 또한 계속해서 좋은 컨텐츠를 생성할 가능성이 높은 사용자로 생각할 수 있다. 이 사용자 신뢰도를 이용하면 검색이나 추천 시 저작자의 신뢰도를 반영하여 컨텐츠의 질을 추정하여 양질의 컨텐츠를 제공해 줄 수 있다.

## 2. 기존 연구

### 2.1 PageRank

PageRank[1]는 웹 문서들 간 하이퍼링크로 연결된 네트워크를 통해 웹 문서의 상대적 중요도를 계산하는 잘 알려진 알고리즘이다. 이 알고리즘을 웹 문서 벡터와 링크 행렬을 이용하여 수식으로 표현하면 다음과 같다.

$$\mathbf{r} = \alpha \cdot \mathbf{T} \cdot \mathbf{r} + (1-\alpha) \cdot \mathbf{d} \tag{1}$$

$$\mathbf{T}(p,q) = \begin{cases} 0 & \text{if } (q,p) \notin \varepsilon \\ 1/w(q) & \text{if } (q,p) \in \varepsilon \end{cases} \tag{2}$$

식 (2)의 행렬 $\mathbf{T}$는 웹 문서의 링크로 이루어지는 네트워크를 나타내며 $w(q)$는 웹 문서 $q$의 out-link의 수를 의미한다. 식 (1)의 $\mathbf{r}(\mathrm{p})$는 웹 문서 $p$의 PageRank 점수를 의미하며 $\mathbf{d}$는 PageRank 알고리즘에서 모든 문서에 대해 같은 값을 갖으며 모든 값의 합이 1인 벡터를 의미한다. 모든 웹 문서에 대해 식 (1)을 $\mathbf{r}$의 값이 수렴할 때 까지 반복적으로 적용하여 얻어진 PageRank 점수를 통해 웹 문서의 중요도를 판단한다.

이 PageRank 알고리즘은 많은 연구에서 응용되어 사용되었다. 대표적으로 TrustRank[2] 알고리즘은 스팸 문서를 구별해내기 위해 신뢰성 있는 문서 셋을 미리 선정하고, 이로부터 웹 문서의 링크 네트워크를 통해 신뢰성 점수를 전파하였다. 이 때 식 (1)의 $\mathbf{d}$ 벡터에 초기 선정된 문서의 신뢰성 점수를 설정하고 나머지 문서는 0으로 설정하는 방법으로 응용하였다.

### 2.2 소셜 네트워크

위의 PageRank 알고리즘을 사용자의 신뢰도를 구하는데 응용하기 위해서는 사용자 간 링크로 연결된 네트워크가 형성되어야 한다. 그러나 웹 문서에서와는 달리 UCC 비디오 서비스 상에서는 사용자가 다른 사용자를 참조하는 명백한 링크는 존재하지 않는다. 하지만 그림 1에 볼 수 있는 것과 같이 UCC 비디오 서비스에서는 한 사용자가 다른 사용자의 컨텐츠를 구독하는 활동, 다른 사용자가 생성한 컨텐츠에 대해 즐겨 찾기에 추가하거나 댓글을 다는 활동 등의 많은 상호작용이 일어난다. 이러한 소셜 활동을 통해 UCC 비디오 서비스를 이용하는 사용자 간 소셜 네트워크를 생성할 수 있다.

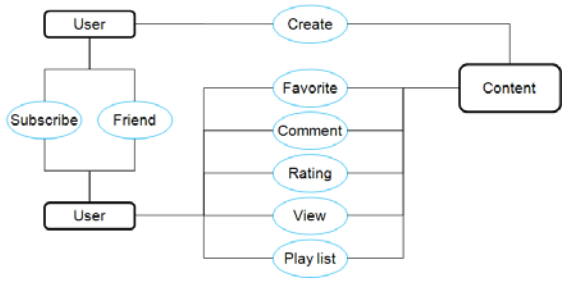


그림 1. UCC 비디오 서비스에서 발생하는 사용자-사용자 간 또는 사용자-컨텐츠 간 상호작용 (YouTube 참고)

## 3. 소셜 네트워크를 통한 사용자 신뢰도

### 3.1 구독 링크를 이용한 사용자 신뢰도

먼저 사용자-사용자 간 소셜 활동이며 특정 사용자의 컨텐츠에 대한 신뢰도를 나타내기에 적당한 구독(subscribe)활동을 링크로 하는 네트워크를 구성하였다. 이 네트워크를 통해 사용자의 신뢰도를 구하기 위해, PageRank 알고리즘을 적용하여 링크 행렬 $\mathbf{T}$를 다음과 같이 정의하였다.

$$\mathbf{T}(p,q) = \begin{cases} 0 & \text{if } (q,p) \notin \varepsilon \\ 1/out_s(q) & \text{if } (q,p) \in \varepsilon \end{cases} \quad (3)$$

위 식에서 $out_s(q)$는 사용자 $q$가 구독하는 모든 사용자의 수를 의미하고 소셜 네트워크상에서는 out-degree를 나타낸다. $\mathbf{T}(p,q)$는 사용자 $q$가 사용자 $p$를 구독할 경우 $q$의 out-degree 역수를 설정하고, 구독하지 않을 경우 0을 설정한다. 이렇게 링크 행렬 $\mathbf{T}$를 정의하고 식 (1)의 $\mathbf{r}$을 모든 사용자의 신뢰도 점수 벡터로 놓고 수렴할 때까지 계산을 반복하여 각 사용자의 신뢰도를 구할 수 있다.

### 3.2 즐겨찾기 링크를 이용한 사용자 신뢰도

사용자-사용자 간 소셜활동 이외에 사용자-컨텐츠 간 소셜활동을 이용하여 네트워크를 구성할 수도 있다. 예를 들어 특정 사용자가 저작한 컨텐츠를 다른 사용자가 즐겨찾기에 추가하는 경우, 신뢰할 수 있는 컨텐츠를 통해 저작자의 신뢰도를 추정할 수 있다. 이 경우 컨텐츠를 매개로 두 사용자 간 즐겨찾기 링크로 연결될 수 있으며, 즐겨찾기 링크와 사용자 노드로 구성되는 소셜 네트워크를 만들 수 있다. 즐겨찾기 링크에 대한 링크 행렬은 다음과 같이 정의할 수 있다.

$$\mathbf{T}(p,q) = \begin{cases} 0 & \text{if } (q,p) \notin \varepsilon \\ n_f(q,p)/out_f(q) & \text{if } (q,p) \in \varepsilon \end{cases} \quad (4)$$

특정 사용자가 저작한 복수개의 컨텐츠에 대해 각각 즐겨찾기를 할 수 있으므로, $n_f(q,p)$를 사용자 $q$가 사용자 $p$의 컨텐츠 중 즐겨찾기에 추가한 컨텐츠의 개수로 정의하였다. 또한 $out_f(q)$는 사용자 $q$가 즐겨찾기에 추가한 모든 컨텐츠의 수를 의미하며 소셜 네트워크상에서 out-degree를 나타낸다. 이와 같이 링크 행렬 $\mathbf{T}$를 정의하고 식 (1)을 수렴할 때까지 반복 계산하여, 즐겨찾기 링크를 통한 사용자 신뢰도를 구할 수 있다.

또한 비슷한 방법으로 댓글을 다는 소셜 활동이나 플레이리스트에 추가하는 소셜 활동을 대상으로 소셜 네트워크를 구성하고 사용자 신뢰도를 구할 수 있다.

### 3.3 통합된 링크를 이용한 사용자 신뢰도

또한 두 종류 이상의 링크를 동시에 이용하여 소셜 네트워크를 구성할 수 있다. 위의 사용자-사용자 간 활동인 구독 활동과 사용자-컨텐츠 간 활동 인 즐겨찾기 활동을 동시에 고려하는 네트워크를 구성하였다. 이 때 각 활동이 사용자 신뢰도에 미치는 영향력이 다를 수 있으므로, 각 링크에 가중치를 $w_s$와 $w_f$로 두어 다음과 같이 링크 행렬을 정의할 수 있다.

$$\mathbf{T}(p,q) = \frac{(w_s \cdot n_s(q,p) + w_f \cdot n_f(q,p))}{\mathrm{W}(q)} \quad (5)$$

$$\mathrm{W}(q) = w_s \cdot out_s(q) + w_f \cdot out_f(q) \quad (6)$$

위 식에서 $n_s(q,p)$는 사용자 $q$가 사용자 $p$를 구독하면 1 그렇지 않으면 0인 함수이고, $\mathrm{W}(q)$는 사용자 $q$에서 나가는 모든 링크에 대한 가중치 합을 나타낸다. 이와 같이 링

크 행렬 $\mathbf{T}$를 정의하면 식 (1)을 통해 구독 링크와 즐겨찾기 링크를 통한 사용자 신뢰도를 구할 수 있다. 또한 비슷한 방법으로 다른 소셜활동에 해당하는 링크를 추가하여, 소셜 네트워크를 구성하고 각 링크에 해당하는 가중치를 주어 사용자 신뢰도를 계산할 수 있다.

### 3.4 비링크 요소의 이용

사용자의 신뢰도를 추정할 때 링크로 표현되는 소셜활동 이외에 링크로 표현되지 않는 값이 영향을 줄 수 있다. 예를 들면 특정 사용자가 저작한 컨텐츠들의 평균 평점 점수는 소셜활동에 해당하지 않지만 사용자의 신뢰도에 영향을 주는 요소가 될 수 있다. 이런 요소를 고려하기 위해서 식 (1)에서 $\mathbf{d}$벡터에 각 사용자에 해당하는 평균 평점 점수를 설정하고 계산을 시작할 수 있다. 이 경우 해당 요소의 중요도에 따라 $\alpha$를 신중히 정해주어야 할 것이다.

## 4. 실험

### 4.1 데이터 수집

본 논문에서는 사용자 생성 비디오 컨텐츠 제공 사이트인 YouTube의 데이터를 대상으로 알고리즘을 적용시켰다. 실험에 필요한 데이터는 YouTube에서 제공하는 API를 통해 크롤링하였으며, 총 데이터는 625,056명의 사용자와 604,903개의 컨텐츠 정보를 모았다. 우선 일정 수의 사용자와 컨텐츠를 기본 데이터 집합으로 시작하여, 소셜 네트워크를 구성하는 각 링크를 통해 추적해나가며 사용자 데이터를 중심으로 크롤링 하였다. YouTube의 데이터가 아주 방대하기 때문에 위의 순서대로 하지 않으면 어느 정도 많은 데이터를 모으더라도 의미있는 소셜 네트워크가 구성되지 않는 경향이 있었다.

### 4.2 결과

위와 같이 수집한 데이터를 대상으로 각각의 소셜네트워크 요소를 적용하여 사용자 신뢰도를 구하였다. 60만명이 넘는 사용자를 대상으로 하였으므로 전체 사용자 중 상위 몇 명의 신뢰도 수치를 비교하면 모두 최대값에 가깝고 큰 차이가 없어서 데이터 비교의 의미가 없었다. 그래서 본 논문에서는 표 1과 같이 YouTube에서 특정 키워드를 검색하였을 때 결과로 나타나는 저작자들의 신뢰도를 비교하였다. 사용자 신뢰도가 높은 저작자의 컨텐츠가 검색 결과에도 대체적으로 상위에 위치되는 모습을 보였으나, 사용자 신뢰도가 높은 저작자의 좋은 컨텐츠가 아주 낮은 순위에 머무르는 경우도 있었다. 랭킹 계산시 사용자 신뢰도를 이용하면 이와 같이 좋은 컨텐츠를 상위로 올릴 수 있을 것이다.

**표 1. YouTube에서 특정 기간 동안 키워드 'ipod'으로 검색한 결과 컨텐츠의 저작자와 그의 신뢰도 (구독 링크와 즐겨찾기 링크를 통합하여 계산)**

| Rank | User ID | User Reputation |
|---|---|---|
| 1 | nigahiga | 0.4558288064 |
| 2 | universalmusicgroup | 0.4237290994 |
| 3 | Daxflame | 0.1009570069 |
| 4 | Blendtec | 0.0976216044 |
| 5 | HouseholdHacker | 0.0959541008 |
| 6 | Blunty3000 | 0.0479920705 |
| 7 | lockergnome | 0.0436237921 |
| 8 | cutewithchris | 0.0392085371 |
| 9 | jimmyrcom | 0.0388136425 |
| 10 | makemagazine | 0.0314857850 |
| 11 | cinemafreaks | 0.0237043214 |
| 12 | RhettandLink | 0.0169548272 |
| 13 | gizmodo | 0.0140463726 |
| 14 | tutvid | 0.0134390355 |
| 15 | thecreativeone | 0.0094674892 |
| 16 | WasteTimeChasingCars | 0.0073195064 |
| 17 | peestandingup | 0.0057146560 |
| 18 | ViewDo | 0.0055285117 |
| 19 | ipodtouchmaster05 | 0.0048007889 |
| 20 | mobuzzES | 0.0047057135 |

## 5. 결론

본 논문에서는 UCC 비디오 공유 서비스에서 검색이나 추천을 통해 사용자들에게 컨텐츠를 제공할 때, 랭킹 알고리즘에서 컨텐츠의 질을 추정할 수 있는 요소로 사용자 신뢰도를 제안하였다. UCC 비디오 서비스를 이용하는 사용자들의 여러 소셜 활동을 통해 소셜 네트워크를 구성하고, 이 네트워크의 링크를 이용하여 사용자의 신뢰도를 계산하는 알고리즘을 정의하였다.

이 알고리즘은 세계 최대의 UCC 비디오 공유 서비스인 YouTube의 데이터를 대상으로 실험하였다. YouTube에서 제공하는 Data API등을 이용하여 데이터를 모아 실험에 이용하였다.

향후에는 전문가의 평가를 통해 실제 사용자의 신뢰도를 정의하고, 이를 이용하여 각 소셜활동 요소가 사용자 신뢰도에 미치는 영향력을 계산하고 이를 해당 요소의 가중치로 반영할 수 있을 것이다. 또한 전문가 평가 리스트를 통해 본 알고리즘의 정확도를 검증 해 볼 수도 있을 것이다.

## 참고문헌

[1] S. Brin and L. Page. "The Anatomy of a large-scale hypertextual Web search engine", Computer Networks and ISDN Systems 30, Elsevier Science B.V., pp 107-117, 1998.

[2] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen, "Combating Web Spam with TrustRank", In Proceedings of the 30th VLDB Conference, Toronto, Canada, pp 576-587, 2004.

[3] Lisa Wiyartani, Yo-Sub Han, Laehyun Kim, "A Ranking Algorithm for User-generated Video Contents based on Social Activities", In Proceedings of Third International Conference on Digital Information Management, London, UK, pp 260-265, 2008.

[4] 조현철, Lisa Wiyartani, 한요섭, 차정원, 김래현, “사용자 신뢰도 기반 UCC 비디오의 랭킹 알고리즘” , 한국정보과학회 제35회 추계학술대회, 서울, 한국, pp 73-74, 2008.