한국시뮬레이션학회 논문지
Vol. 19, No. 1, pp. 83-90 (2010. 3)

콘텐츠 명성 및 사용자 명성 평가를 이용한 UCC 검색 품질 개선

배원식[1] · 차정원[1†]

Improving the Performance of the User Creative Contents Retrieval Using Content Reputation and User Reputation

Won Sik Bae · Jeong Won Cha

ABSTRACT

We describe a novel method for improving the performance of the UCC retrieval using content reputation and user reputation. The UCC retrieval is a part of the information retrieval. The goal of the information retrieval system finds documents what users want, so the goal of the UCC retrieval system tries to find UCCs themselves instead of documents. Unlike the document, the UCC has not enough textual information. Therefore, we try to use the content reputation and the user reputation based on non-textual information to gain improved retrieval performance. We evaluate content reputation using the information of the UCC itself and social activities between users related with UCCs. We evaluate user reputation using individual social activities between users or users and UCCs. We build a network with users and UCCs from social activities, and then we can get the user reputation from the network by graph algorithms. We collect the information of users and UCCs from YouTube and implement two systems using content reputation and user reputation. And then we compare two systems. From the experiment results, we can see that the system using content reputation outperforms than the system using user reputation. This result is expected to use the UCC retrieval in the feature.

Key words : UCC, Information Retrieval, Social Activity, Content Reputation, User Reputation

요 약

본 논문에서는 콘텐츠 명성 및 사용자 명성 평가를 통해 신뢰성 높은 UCC 검색을 가능하게 하는 방법에 대해 기술한다. 기존 정보검색과 달리 UCC에서는 얻을 수 있는 텍스트 정보가 한정적이기 때문에 텍스트 외적인 정보의 사용이 필요하다. 콘텐츠 명성과 사용자 명성은 비텍스트 정보를 이용하여 평가되는데, 평가된 명성을 자질로 사용하여 UCC 검색을 수행하면 기존 검색 방법보다 향상된 검색 성능을 기대할 수 있다. 콘텐츠 명성은 영상 자체 정보와 영상과 관련된 소셜활동 정보로부터 콘텐츠의 명성, 즉 가치를 평가한다. 또한 사용자 명성은 콘텐츠와 사용자, 사용자와 사용자 사이의 소셜활동 하나하나에 주목하여 네트워크를 구축하여 사용자의 명성을 평가한다. 각각의 명성을 평가하여 UCC 검색에 사용하는 두 개의 시스템을 구현하고, 유튜브로부터 수집한 UCC와 사용자 정보를 이용하여 두 시스템의 비교 실험을 수행하였다. 실험 결과, 콘텐츠 명성을 활용한 시스템에서 조금 더 높은 사용자의 동의를 이끌어 낼 수 있었으며, 이 결과는 향후 UCC 검색에 활용할 수 있을 것이라 기대된다.

주요어 : UCC, 정보검색, 소셜활동, 콘텐츠 명성, 사용자 명성

* 본 연구는 지식경제부 및 정보통신산업진흥원(구 정보통신연구진흥원)의 IT핵심기술개발사업의 일환으로 수행하였음(과제관리번호: 2008-S-024-01, Rich UCC 기술개발).
2009년 10월 12일 접수, 2010년 1월 15일 채택
[1] 창원대학교 컴퓨터공학과
주 저 자 : 배원식
교신저자 : 차정원
E-mail; jcha@changwon.ac.kr

1. 서 론

UCC(User Creative Contents)는 사용자가 직접 제작한 콘텐츠를 말한다. 인터넷과 디지털카메라, 휴대전화 등의 정보통신 매체들이 활발하게 보급됨에 따라 널리 확산되었으며, UCC 서비스를 제공하는 포털 사이트들이 많이 생겨났다. 대표적인 사이트로는 유튜브[1], 판도라TV[2],

곰TV[3], 아프리카[4] 등이 있다. 수많은 UCC가 공개되고, 서비스에 이용되면서 사용자뿐만 아니라 UCC 서비스를 제공하는 업체들도 높은 신뢰도의 UCC 검색을 필요로 하고 있다. 따라서 UCC 검색은 사용자의 질의를 분석하여 원하는 문서를 찾는 정보검색(Information Retrieval)의 한 분야로 새롭게 자리 잡아가고 있다.

정보검색 분야에서 많이 사용되는 방법 중 하나는 단어 통계량 정보를 기반으로 하는 TF-IDF 방법(G. Salton 등, 1998)이다. 그러나 이 방법은 텍스트 형태의 문서 검색에는 적합하지만 UCC 검색에는 적합하지 못하다. 왜냐하면 UCC는 영상에 검색에 중요한 정보가 담기므로 상대적으로 텍스트는 부가 정보 정도만 기술하는 제한적인 용도로 사용되어 UCC의 특징을 잘 반영하지 못하기 때문이다. 따라서 UCC 검색에서는 텍스트 외적인 자질, 즉 비텍스트 자질을 사용할 필요가 있다. 다행스럽게도 UCC에는 조회수나 스크랩 횟수, 평점과 같은 비텍스트 자질이 존재한다. 이 자질들은 수많은 사용자들에 의해 평가된 것이므로 UCC의 가치를 평가하는데 좋은 도구가 될 수 있다. 또한 사용자와 사용자, 사용자와 UCC 사이의 소셜활동으로부터 네트워크를 구축할 수 있는데, 그래프 알고리즘을 통해 구축한 네트워크에서 UCC나 사용자의 가치를 평가할 수 있다. 본 논문에서는 이러한 비텍스트 자질로부터 콘텐츠 명성과 사용자 명성을 평가하고, 평가한 명성을 UCC 검색에 사용하여 검색의 신뢰성을 높일 수 있는 방법을 제안한다. 명성(Reputation)은 "세상에 널리 퍼져 평판이 높은 이름"이라는 사전적 의미를 담고 있다. UCC 검색에서 "세상"은 콘텐츠(UCC)와 사용자 사이에 구축되는 가상의 공간이며, 이 공간에서 콘텐츠와 사용자는 각각의 명성을 갖는다. 따라서 명성을 평가하여 UCC 검색에 활용하면 검색의 신뢰성을 향상시킬 수 있을 것이라 판단하였다. 본 논문에서는 다음과 같은 방법으로 콘텐츠와 사용자의 명성을 평가한다. 먼저 콘텐츠의 명성 평가에는 조회나 평점부여와 같은 영상 관련 소셜활동 정보와 화질 등과 같이 영상 자체적으로 얻을 수 있는 정보를 사용한다. 사용자의 명성 평가에는 영상과 관련된 사용자들의 소셜활동 정보와 사용자 간의 소셜활동 정보로부터 네트워크를 구축하고 사용한다. 이 네트워크는 그래프 형태로 표현이 가능하므로 PageRank 알고리즘(S. Brin 등, 1998) 등의 그래프 알고리즘을 적용하여 UCC나 사용자의 중요도, 즉 명성을 평가할 수 있다. 콘텐츠 명성과 사용자 명성을 평가하기 위하여 소셜활동 정보를 각각 사용한다. 하지만 콘텐츠 명성은 사용자들의 UCC에 대한 전체적인 소셜활동 정보에 주목하여 평가하는데 반해 사용자 명성은 사용자의 소셜활동 하나하나에 주목하여 평가하기 때문에 두 명성은 서로 다른 특성을 갖는다. 따라서 본 논문에서는 각각의 명성 평가 방법을 이용하여 UCC를 검색하는 두 개의 시스템을 구현하고, 유튜브에서 수집한 데이터를 사용하여 실험을 수행하였다.

본 논문의 구성은 다음과 같다. 2장에서는 비텍스트 자질을 이용한 기존 연구들에 대해서 살펴보고, 3장에서는 콘텐츠와 사용자의 명성을 평가하는 방법과 평가된 명성을 사용하는 UCC 검색 시스템에 대하여 자세히 설명한다. 4장에서는 실험에 대한 전반적인 내용과 실험 결과를 정리하고, 끝으로 5장에서는 결론과 향후 과제를 다룬다.

2. 관련연구

서론에서 언급한 것처럼 UCC 검색에서는 텍스트로 부터 얻을 수 있는 정보가 제한적인 것이 문제가 된다. 따라서 전통적으로 사용해 온 텍스트 자질과는 비텍스트 자질과 같은 자질의 사용을 고려해볼 필요가 있다. 비텍스트 자질을 UCC 검색에 적용한 연구는 없었지만 정보검색이나 질의응답 분야에서 답변 문서의 품질을 평가하기 위하여 진행된 연구를 찾을 수 있었다. 본 장에서는 그 연구들에 대해 간략히 정리하고자 한다.

전지원 등은 총 13개의 비텍스트 자질을 추출하여 답변 문서의 품질 평가에 사용하였다(J. Jeon 등, 2006). 13개 자질에는 응답자의 답변 이력, 활동성, 활동 범주와 같은 응답자 관련 자질과 답변의 길이와 같은 문서 자체 자질, 추천수나 조회수와 같이 시간의 흐름에 따라 얻어지는 통계적 자질 등이 포함된다. 이렇게 추출된 자질과 답변의 평가 정보 사이의 상관 계수(Correlation Coefficient)를 계산하여 자질을 분석하고, KDE(Kernel Density Estimation) 방법(J. Hwang 등, 2006)을 이용하여 자질 변환(Feature Conversion)하고, 최대 엔트로피 모델(A. Berger 등, 1996)을 사용하여 답변 문서의 품질을 평가하였다.

이정태 등(2007)은 비텍스트 자질을 대신하여 명성 자질을 지식검색 문서의 품질 평가에 사용하였다. 비텍스트 자질이 문서의 품질 평가에 유효한 정보임에는 틀림없지만 새로 작성된 문서는 데이터 부족 문제(Data Sparseness)가 발생할 수 있다는 점을 문제 삼고 있다. 이러한 데이터

1) YouTube, http://www.youtube.com
2) http://www.pandora.tv
3) http://www.gomtv.com
4) http://www.afreeca,com

부족 문제에 대응하기 위하여 문서 자체로부터 문서의 신뢰성을 반영할 수 있는 '명성 자질'이라고 명명된 9개 정도의 자질을 사용하였다. 명성 자질에는 연결어(접속사)의 출현 비율, 불확실한 추정어의 출현 비율, 광고성 단어의 출현 비율, 이모티콘의 출현 비율, 웹사이트 주소 출현 여부 등과 같은 문서로부터 직접적으로 얻을 수 있는 자질이 포함되며, 사전기반 자질이 큰 비중을 차지한다. 명성 자질로부터 A. Berger 등(1996)과 같이 최대 엔트로피 모델을 사용하여 문서의 품질을 평가한다. 명성 자질은 지식검색 문서와 같이 지식 전달을 목적으로 작성된 문서에서는 유효하지만, UCC 검색에서와 같이 텍스트가 추가적인 정보를 기술하는 용도로 주로 사용되면 효용성이 낮아질 수밖에 없다.

이현우 등(2009)은 동시출현 자질과 사용자의 활동성에 기인한 사용자 명성 평가를 통해 지식검색 문서의 품질을 평가하고자 하였다. 이정태 등(2007)과 마찬가지로 비텍스트 자질을 사용하는 것은 데이터 부족 문제를 야기하며, 비텍스트 자질을 지속적으로 관리하고 보강하는 작업이 뒤따라야만 한다는 것을 문제점으로 꼽고 있다. 이러한 문제점을 해결하기 위하여 사용자의 활동 내역으로부터 그래프로 구성하고, PageRank 알고리즘을 사용하여 해당 사용자의 명성을 평가하여 문서의 품질 평가에 사용하였다. 또한 단순히 답변만 다수 작성하는 사용자의 명성이 지나치게 높아지는 것을 방지하기 위하여 질문과 답변과의 유사성을 계산하기 위한 동시출현 자질도 사용하였다. "네이버 지식iN[5]"에서 수집한 데이터를 이용한 실험에서 "네이버 지식iN" 서비스에서 제공하는 사용자 답변 선택률 순위와 비슷한 결과를 보여주었다. 이 결과는 수집한 데이터를 제안한 방법 외에 특별한 처리를 하지 않고 얻어졌는데, 본 논문에서 다루고 있는 UCC 검색 및 다른 인터넷 서비스에서도 충분한 가능성이 있다는 것을 말해주는 것이라 할 수 있다.

3. 콘텐츠 명성 및 사용자 명성 평가와 명성을 이용한 UCC 검색 시스템

본 장에서는 콘텐츠 명성과 사용자 명성을 평가하는 방법 및 평가된 명성을 UCC 검색에 활용하는 방법에 대하여 자세히 기술한다.

3.1 콘텐츠 명성 평가

3.1.1 자질 정의

콘텐츠 명성 평가 알고리즘에 사용하는 자질은 크게 영상 자체 정보와 영상 관련 소셜활동 정보로 구분할 수 있는데, 각각의 자질에 대한 구체적인 설명은 다음과 같다.

- **영상 자체 정보 :** 영상 자체 정보는 사용자와는 상관없이 UCC 영상 자체가 갖고 있는 자질이다. UCC는 최근에 올라온 영상이 중요도와 이슈가 될 가능성이 높으며, 시간이 흐를수록 중요도가 낮아지는 특징을 가지고 있다. 또한 사용자들은 일반적으로 같은 영상일 경우에는 화질이 더 좋은 영상을 선호하는 경향이 있다. 따라서 영상 자체 정보에는 UCC가 업로드 후 경과한 시간(T), UCC 영상의 화질(Q)이 포함된다.
- **영상 관련 소셜활동 정보 :** 영상 관련 소셜활동 정보는 사용자들의 UCC와 관련된 소셜활동으로부터 수집된 통계 정보 기반의 자질이다. 소셜활동에는 UCC를 본 사용자의 수(V), 스크랩한 사용자의 수(S), 평점을 부여한 사용자의 수(R), 사용자들이 부여한 평점의 평균(AR)이 포함된다. 댓글과 관련한 통계 정보도 존재하지만 의미 없는 댓글이나 악성 댓글과 같은 잡음 정보들이 다수 존재하므로 댓글은 자질로 사용하지 않았다.

3.1.2 영상 관련 소셜활동 정보의 비례축소와 평준화

영상 관련 소셜활동 정보의 세부 정보들에 대한 통계량은 표 1과 같다. 우선 조회수가 스크랩 횟수나 평점부여 횟수보다 절대적인 값의 크기가 크다는 것을 알 수 있다. 따라서 각 자질들을 함께 조합하기 위해 비례 축소(Scaling)를 수행해야 한다.

그림 1은 조회수가 많은 순서대로 상위 1만 개의 비디오에 대한 조회수의 분포를 나타내는 그래프이다. 그래프의 가로축은 비디오 아이디를 나타내며, 비디오를 조회수가 많은 순서대로 왼쪽부터 차례로 배치하였다. 세로축은 특정 비디오에 대한 조회수를 나타낸다.

그래프를 살펴보면 비디오 아이디가 증가할수록 조회

표 1. 영상 관련 소셜활동 정보의 통계 정보

구분	최대	평균	표준편차
조회	125,255,461	254,183	1,223,701
스크랩	779,836	996	5,160
평점부여	585,145	601	3,192

5) http://kin.naver.com/

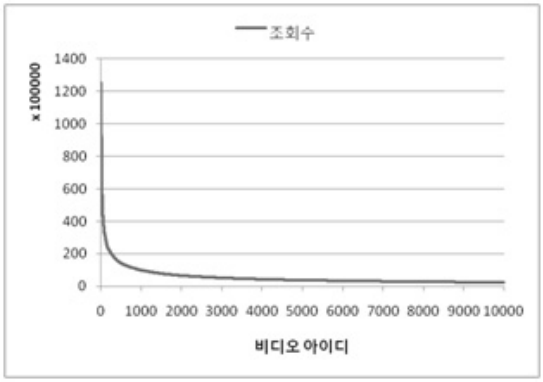

그림 1. 비디오에 대한 조회수 분포

수가 급격하게 떨어지는 것을 확인할 수 있다. 전체 비디오에 대한 그래프로 확장을 하거나 스크랩 횟수, 평점부여 횟수에 대한 그래프도 마찬가지 형태를 갖는다. 따라서 이들 통계량을 그대로 사용하면, 높은 조회수를 갖는 비디오들과 나머지 비디오들 사이의 편차가 심해져 자질로서의 역할을 제대로 수행할 수 없게 된다. 따라서 자질로 사용하기 위하여 표준화(Normalization)를 수행해야 한다.

본 논문에서는 식 (1)의 Sigmoid 함수를 사용하여 비례 축소와 평준화를 동시에 수행한다. Sigmoid 함수를 적용하면 각 자질의 값은 0과 1 사이의 값으로 얻어진다. 또한 조회수가 높은 상위 비디오들이 대부분의 조회수를 차지하여 나머지 대다수의 비디오들이 전체적으로 낮은 값을 가져 자질로서의 의미를 상실하는 것을 피할 수 있다.

$$P(x) = \frac{1}{1+e^{-ax}} \tag{1}$$

식 (1)에서 α는 x의 크기에 따른 확률 P(x)의 증가 폭을 결정하는 인자이다. Sigmoid 함수의 기본형에서 α의 값은 1이며, x가 4 이상일 때 P(x)는 거의 1의 값을 갖는다. 표 1과 같이 조회수, 스크랩 횟수, 평점부여 횟수는 서로 다른 통계량을 갖기 때문에 각각 다른 α값을 사용하여 평준화를 수행한다. α값은 조회수에서는 0.000005, 스크랩과 평점부여 횟수에서는 0.00005의 값을 사용한다.

3.1.3 콘텐츠 명성 평가 모델

식 (2)는 콘텐츠 명성을 평가하기 위한 모델을 나타낸 것이다. 모델에는 선행 단계에서 비례축소와 표준화를 통해 얻어진 값을 자질로 사용한다. 콘텐츠 명성은 영상 자체 정보와 영상 관련 소셜활동 정보를 α:β의 비율로 선형 조합(Linear Combination)하여 평가한다. 콘텐츠 명성 평가 모델로부터 얻어진 명성값도 0과 1 사이의 값을 갖는다.

$$CR = \frac{\alpha(T+Q)+\beta(\gamma V+S+R+AR)}{2\alpha+(3+\gamma)\beta} \tag{2}$$

식 (2)에서 가중치 α와 β의 값, 즉 영상 자체 정보와 영상 관련 소셜활동 정보의 조합 비율은 1: $\sqrt{2}$ 이다. 상대적으로 소셜활동 정보가 영상 자체 정보보다 명성을 평가하는데 중요한 역할을 할 것이라는 사실에는 이견이 없을 것이다. 그러나 영상 자체 정보 또한 무시할 수 없는 자질이므로 두 가중치가 너무 차이가 나지 않도록 가중치를 조정하였다. 가중치 조정에 대한 내용은 4장에서 조금 더 자세히 기술하도록 하겠다.

가중치 γ는 조회수 자질(V)에 대한 가중치로써 10의 값을 갖는다. 스크랩이나 평점부여는 사용자의 성향에 따라 잘 행해지지 않는 경우도 있지만, 조회는 사용자의 성향과는 관계없이 항상 행해지므로 명성을 평가하는데 사용할 중요한 척도가 될 수 있다. 실제로 국내외 UCC 서비스를 비교해서 살펴보면, 국외에 비해 국내 서비스에서는 평점부여 수나 댓글 수, 스크랩 수가 현저히 적다는 사실을 쉽게 확인할 수 있다.

3.2 사용자 명성 평가

사용자 명성 평가는 본 논문의 선행 연구 격인 Y. Han 등(2009)의 연구에서 제안한 모델을 따른다. 앞에서 언급한 것과 같이 사용자 명성은 사용자와 사용자, 사용자와 콘텐츠 간의 소셜활동으로부터 네트워크를 구축하여 평가에 사용한다. 그림 2는 유튜브로부터 구축된 네트워크 중 일부를 보여주는 그래프이다. 원과 화살표는 각각 그래프의 노드(Node)와 간선(Edge)을 나타낸다.

사용자와 콘텐츠 모두가 그래프의 노드가 될 수 있으며, 노드의 크기는 해당 노드의 가치를 나타낸다. 자신으로 향하는 간선이 많으면 가치가 높아지고, 자신 노드에서 다른 노드로 향하는 노드가 있으면 자신의 가치를 나누어준다. 여기서 말하는 가치가 곧 명성이다. 간선은 크게 네 종류의 소셜활동에 의해 연결된다. 첫 번째는 사용자가 콘텐츠를 업로드하는 활동이다. 사용자가 콘텐츠를 업로드하면 사용자에서 콘텐츠로 향하는 간선이 연결되고, 반대 방향의 간선도 연결된다. 반대 방향의 간선은 가치가 높은 콘텐츠를 올린 사람이 명성이 높아지도록 만들어주기 위한 피드백(Feedback) 역할을 한다. 두 번째는

특정 사용자가 다른 사용자를 구독(Subscription, 관심 사용자 등록)하는 활동이다. 이 경우에는 구독을 신청한 사용자에서 수락한 사용자 쪽으로 간선이 연결된다. 세 번째는 특정 사용자가 특정 콘텐츠를 즐겨찾기하는 활동이다. 두 번째와 마찬가지로 즐겨찾기를 신청한 사용자에서 대상이 되는 콘텐츠 쪽으로 간선이 연결된다. 끝으로 특정 사용자가 특정 콘텐츠에 댓글을 다는 활동이다. 마찬가지로 댓글을 단 사용자에서 댓글이 달리는 콘텐츠 쪽으로 간선이 연결된다.

콘텐츠 명성은 위와 같이 구축된 그래프로부터 PageRank 알고리즘을 응용한 식 (3)에 의해 평가한다.

$$UR(U_i) = d + (1-d) \times \sum_{T_j \in In(U_i)} \frac{w(T_j) \times UR(T_j)}{|Out(T_j)|} \qquad (3)$$

식 (3)에서 d는 PageRank 알고리즘의 제동계수(Damping Factor)와 동일하다. 이 알고리즘은 명성값이 일정한 값으로 수렴되기 전까지 반복(Iteration)하여 명성을 평가한다. U_i는 사용자 명성을 평가하려는 노드이고, T_j는 U_i의 사용자 명성을 평가하는데 영향을 미치는 노드이다. $In(U_i)$는 노드 i로 향하는 간선을 가진 노드의 집합을 의미하고, $|Out(T_j)|$는 노드 j에서 다른 노드로 향하는 간선을 가진 노드의 집합의 크기를 의미한다. $w(T_j)$는 노드 j의 가중치인데, 간선의 종류에 따라 다른 값을 갖는다. 구독, 업로드, 즐겨찾기의 순으로 0.35, 0.3, 0.2의 가중치를 가지며, 그 외의 간선은 0.15의 가중치를 갖는다. 콘텐츠도 노드이므로 별도로 명성이 평가되지만 알고리즘에 의해 콘텐츠 명성이 사용자 명성에 영향을 미치므로 콘텐츠 명성을 따로 사용하지는 않는다.

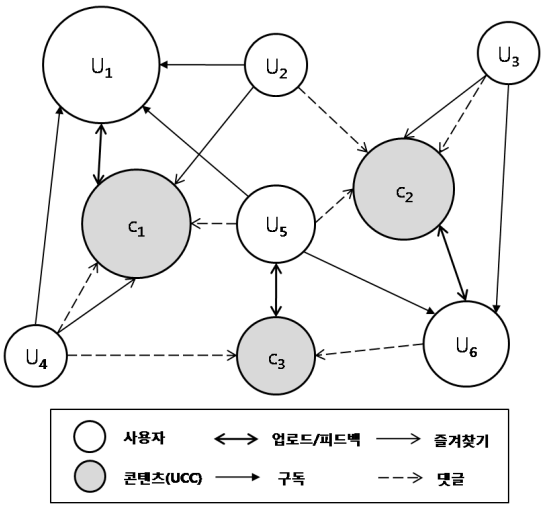

그림 2. 사용자 명성 평가를 위해 구축한 그래프의 일부

3.3 명성 자질을 이용한 UCC 검색 시스템

본 논문에서 제안하는 명성을 이용한 UCC 검색 시스템은 정보검색 분야에서 잘 알려져 있는 키워드 기반의 Okapi (BM25) 모델(S. E. Robertson 등, 1995)을 기반으로 구현하였다. 기본적으로 UCC는 제목, 태그, 영상, 본문, 댓글로 구성되어 있는데, 제목, 태그, 본문에서 나타난 키워드만 색인(Indexing)에 사용하였다. 불용어(Stop-words)는 제거하였고, 스테밍(Stemming)은 수행하지 않았다. 키워드는 제목, 태그, 본문 중에서 나타난 위치에 따라 다른 가중치를 부여하여 Okapi 모델에 사용된다. 가중치는 제목, 태그, 본문에 각각 10, 5, 1의 값을 부여하였다. 태그는 경우에 따라 제목보다 더 중요하게 사용될 수 있지만 사용자의 불순한 의도에 따라 UCC와 관계없는 키워드가 무분별하게 사용될 가능성이 존재한다. 따라서 제목보다는 가중치를 낮게 부여하였다. 콘텐츠 명성과 사용자 명성은 일종의 가중치로써 Okapi 모델과 함께 검색 결과에 영향을 미친다.

4. 실 험

본 장에서는 본 논문의 성능을 평가하기 위한 실험 데이터에 대하여 살펴보고, 실험 방법 및 성능 평가 방법, 실험 결과에 대해 설명한다.

4.1 실험 데이터

본 논문에서는 유튜브로부터 비디오와 사용자 정보를 수집하여 실험 데이터로 사용하였다. ‘ipod’을 씨앗 단어(Seed Word)로 하여 검색된 비디오로부터 관련된 비디오와 사용자를 연쇄적으로 수집하였다. 수집된 총 비디오와 사용자의 수는 표 2와 같다.

4.2 성능 평가 방법

콘텐츠 명성과 사용자 명성 평가의 신뢰도는 직접적으로 평가하기가 힘들다. 따라서 각각의 명성을 이용한 검

표 2. 수집된 비디오 수와 사용자 수

구분	수집량
비디오	604,903
사용자	625,066

색 시스템의 성능을 평가하여 신뢰도를 평가하는 것을 대체하고자 한다. 검색 시스템의 성능 평가는 시스템의 검색 순위를 사용자가 얼마나 동의할 수 있는지를 평가하는 방식을 사용하였다. 이것은 일반적으로 정보 검색이나 질의응답 시스템에서 MRR(Mean Reciprocal Rank) 방법을 사용하여 질의와 관련 있는 1위 문서나 답변을 시스템이 얼마나 잘 찾아내는지를 평가의 기준으로 삼는 것과는 다소 차이가 있다. UCC 검색은 사용자가 원하는 정보를 얻고자 하는 목적 외에도 즐거움을 찾기 위해 검색을 하게 되는 경우가 많다. 즐거움은 사용자의 성향에 따라 기준이 서로 다르기 때문에 객관적으로 순위를 매기기가 어렵다. 따라서 본 논문에서는 시스템이 객관적인 순위를 잘 찾아내는지를 평가하기 보다는 시스템이 내놓은 순위에 대하여 사람이 얼마나 동의할 수 있는지를 평가하는 것으로 성능 평가를 수행하였다.

실험을 진행한 방법은 다음과 같다. 먼저 콘텐츠 명성과 사용자 명성을 이용한 두 개의 시스템의 검색 결과 중 상위 5등까지의 비디오를 사용자에게 보여주었다. 그리고 사용자의 주관에 따라 1등부터 5등까지의 순위를 다시 매기도록 하였다. 질의어로는 수집된 비디오에서 불용어(Stop-words)를 제외하고 남은 빈도가 높은 단어 중에서 사용자가 순위를 매기기 좋을 것 같은 단어를 선택하였다. 표 3과 같은 10개의 질의어를 사용하였으며, 7명의 사용자가 실험에 참여하였다. 사용자들이 평가한 순위를 종합하여 식 (4)를 적용하여 시스템의 성능을 평가하였다.

표 3. 가중치 최적화를 위한 7가지 실험 정의

실험	정의
1	영상 관련 소셜활동 정보 자질만 사용 (α=0, β=1)
2	영상 자체 정보 자질만 사용 (α=1, β=0)
3	두 자질을 1:1로 조합 (α=β=1)
4	영상 관련 소셜활동 정보 자질에 $\sqrt{2}$ 배 가중치 부여(α=1, β=$\sqrt{2}$)
5	영상 자체 정보 자질에 $\sqrt{2}$ 배 가중치 부여 (α=$\sqrt{2}$, β=1)
6	영상 관련 소셜활동 정보 자질에 $\sqrt{2}$ 배 가중치 부여(α=1, β=2)
7	영상 자체 정보 자질에 2배 가중치 부여 (α=2, β=1)

$$k = \frac{1}{Q}\sum_{j=1}^{Q}\frac{1}{Z}\sum_{i=1}^{n}(n+1-i)\frac{1}{rank_i} \quad (4)$$

식 (4)에서 Q는 질의의 수, n은 의미 있는 순위의 수이며, Z는 평준화 인자(Normalization Factor)이다. 그리고 ranki는 시스템의 i번째 순위의 비디오를 사용자가 다시 평가하였을 때의 순위이다.

4.3 콘텐츠 명성 평가 모델의 가중치 조정

3.2절에서 살펴본 것과 같이 콘텐츠 명성 평가 모델에는 영상 자체 정보 자질과 영상 관련 소셜활동 정보 자질을 조합하기 위하여 α, β의 가중치를 사용한다. 본 절에서는 실험을 통해 가중치 α, β의 값에 따라 UCC 검색 성능이 어떻게 변화하는지 확인하고, 최적화된 가중치 값을 얻고자 한다. 실험은 표 3과 같이 총 7가지로 형태로 진행하였는데 영상 자체 정보와 영상 관련 소셜활동 정보 자질을 각각 사용하는 실험에서부터 각각의 자질을 1:2의 비율로 조합하는 것까지 다양하게 수행하였다.

표 4는 각각의 실험에 대한 UCC 검색 성능의 변화를 정리한 것이다.

표 4를 살펴보면 영상 관련 소셜활동 정보 자질에 $\sqrt{2}$의 가중치를 부여한 실험 4에서 0.817로 가장 높은 성능을 나타내는 것을 확인할 수 있다. 영상 관련 소셜활동 정보 자질만 사용한 실험 1에서 0.815로 최고 성능에 근접한 성능을 보여주고 있다. 또한, 영상 자체 정보 자질을 영상 관련 소셜활동 정보 자질보다 높은 가중치로 조합한 실험 5와 7에서 각각 0.759, 0.770의 가장 낮은 성능을 보여주고 있다. 이 결과는 영상 자체 정보 자질보다는 영상 관련 소셜정보가 UCC 검색의 신뢰도에 영향을 더 많이 미치는 자질임을 말해주는 것이다. 그렇지만 영상 자체 정보 자질만 사용한 실험 1에서도 0.779로 어느 수준의 성능을 보여주고 있으며, 실험 3과 4를 통해 두 자질의

표 4. 가중치에 따른 UCC 검색 성능 변화

실험	성능
1	0.815
2	0.779
3	0.783
4	**0.817**
5	0.759
6	0.773
7	0.770

표 5. UCC 검색 성능표

질의어	CR	UR
animation	0.866	**0.893**
naruto	**0.833**	0.772
myspace	0.818	**0.860**
guitar	**0.897**	0.678
lyrics	0.918	**0.990**
ipod	**0.807**	0.734
apple	**0.820**	0.734
google	0.634	**0.912**
iphone	0.600	**0.632**
bush	**0.981**	0.875
전체	**0.817**	0.808

조합 비율에 따라 성능이 향상되는 것을 확인할 수 있다. 이 결과로부터 영상 자체 정보 자질도 영상 관련 소셜활동 정보와 함께 UCC 검색의 신뢰도 향상에 영향을 주는 무시할 수 없는 자질임을 확인할 수 있다.

4.4 실험 결과

표 5는 콘텐츠 명성(CR)과 사용자 명성(UR)을 기반으로 구현된 UCC 검색 시스템의 성능을 정리한 표이다. 10개의 질의어에 대한 성능과 전체를 종합한 성능을 함께 실었다. 각 질의어 중 굵은 글씨로 표시한 것은 두 시스템 중 더 높은 성능을 나타냈다는 것을 의미한다.

콘텐츠 명성 기반 시스템이 사용자 명성 기반 시스템보다 근소하게 높은 성능을 보이고 있다. 물론 모든 질의어에 대해 콘텐츠 명성을 사용한 시스템이 높은 성능을 보이는 것은 아니기 때문에 질의어를 변경한다면 다른 결과가 얻어질 수 있을 것이다. 하지만 'ipod'을 기초로 UCC를 수집하였으므로 수집된 UCC 중에는 'ipod'과 관련된 것이 많다. 따라서 다른 질의어보다도 'ipod'을 질의어로 사용한 실험의 신뢰도가 상대적으로 높을 수 있는데, 해당 경우 콘텐츠 명성을 사용한 시스템 쪽의 성능이 다소 높은 것을 확인할 수 있다. 이 결과는 상대적으로 복잡한 과정을 통해 얻어지는 사용자 명성과 비교했을 때 보다 단순한 콘텐츠 명성을 사용하는 것 또한 나쁘지 않다는 것을 말해주는 결과이다.

5. 결론 및 향후 과제

본 논문에서는 비텍스트 자질을 사용하여 콘텐츠 명성 및 사용자 명성을 평가하고, 평가한 명성을 UCC 검색에 사용하여 검색의 신뢰성을 향상시키는 방법을 제안하였다. 사용자 명성은 UCC 세상에서 발생하는 소셜활동으로부터 구축한 네트워크로부터 평가한다. 따라서 단순한 빈도수에 의존하지 않고 신뢰도가 높은 명성값을 얻을 수 있다는 장점이 있다. 또한 사용자의 명성이 콘텐츠에 영향을 미치므로 새로 등록된 콘텐츠에도 유연하게 대응할 수 있다는 장점이 있다. 그러나 네트워크 구축과 명성 계산에 많은 비용이 요구되기 때문에, 빠르게 변화하는 UCC에 대응할 수 있도록 빈번한 명성값의 업데이트가 어렵다는 문제점이 있다. 콘텐츠 명성은 빈도수에 영향을 많이 받으므로 일정 수 이상의 빈도수가 누적이 되어야 자질로서의 신뢰성을 보장받을 수 있다는 약점이 존재한다. 하지만 빈번한 업데이트가 가능하므로 빠르게 변화하는 UCC의 흐름에 대응할 수 있으며, 전체 누적 빈도수뿐만 아니라 시간에 따른 빈도수 활용을 통해 사용자 명성에 못지않은 신뢰도를 기대할 수 있다. 또한 실험 결과에서도 콘텐츠 명성 사용한 검색이 근소하지만 높은 성능을 보이고 있으므로, 콘텐츠 명성 평가를 사용하는 UCC 검색의 가능성을 확인할 수 있는 결과이다. 두 명성 평가를 사용하는 UCC 검색 시스템에 대한 연구가 계속 진행 중이며, 콘텐츠 명성 평가를 사용하는 UCC 검색 시스템은 베타 서비스를 통해 실제 사용자들의 평가를 받고 있다.

본 논문의 UCC 검색 시스템은 명성 평가를 이용하여 UCC 검색을 수행하고 있지만 기본적으로 키워드 기반의 검색을 수행하기 때문에 검색 결과도 키워드에 의존적이라는 문제점을 안고 있다. 특히 UCC에 붙어 있는 태그의 경우에는 작성에 특별한 제약 사항이 없으므로 작성자의 의도에 따라 무분별하게 붙여지는 경우가 많아 검색 결과에 나쁜 영향을 미치는 경우가 많았다. 차후 태그 정련 방법을 적용하여 검색에 활용하면 현재보다 개선된 품질의 UCC 검색이 가능할 것으로 기대된다. 또한 지금과 같이 평가한 명성값을 Okapi 모델의 가중치로 사용하는데 그치지 않고, 조금 더 비중 있게 활용할 수 있는 방법을 연구하여 UCC 검색의 품질을 더욱 개선할 수 있도록 할 것이다.

참 고 문 헌

1. 이정태, 송영인, 임해창, "명성 자질을 이용한 지식검색 문서의 품질 평가," 제19회 한글 및 한국어 정보처리 학술대회(HCLT 2007) 논문집, pp. 62-67, 2007년.
2. 이현우, 한요섭, 김래현, 차정원, "동시출현 자질과 집단 지

성을 이용한 지식검색 문서 사용자 명성 평가," 한국인지과학회논문지:인지과학, 19(5), pp. 459-476, 2009년.
3. A. Berger, S. D. Pietra, and V. D. Pietra, "A Maximum Entropy Approach to Natural Language Processing," Computation Linguistics, vol. 22, no. 1, pp. 39-71, 1996.
4. G. Salton and C. Buckley, "Term-weighting Approach in Automatic Text Retrieval," Information Processing & Management, vol. 24, no. 5, pp. 513-523, 1998.
5. J. Hwang, S. Lay and A. Lippman, "Nonparametric Multivariate Density Estimation: A Comparative Study," IEEE Transactions of Signal Processing, vol. 42, no. 10, pp. 2795-2810, 1994.
6. J. Jeon, W. B. Croft, J. H. Lee., and S. Par, "A Framework to Predict the Quality of Answers with Non-Textual Features," Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 228-265, 2006.
7. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Computer Networks and ISDN Systems, vol. 30, pp. 107-117, 1998.
8. S. E. Robertson, S. Walker, S. Jones, M. M. Hancook-beauliew, and M. Gatfrod, "Okapi at TREC-3," Proceedings of 3rd Text Retrieval Conference(TREC-3), pp. 109-126, 1995.
9. Y. Han, L. Kim, and J. Cha, "Evaluation of User Reputation on YouTube," Proceedings of the 3rd International Conference on Online Communities and Social Computing: Held as Part of HCI International 2009, vol. 12, pp. 346-353, 2009.

배 원 식 (wonsigi529@changwon.ac.kr)

2007 창원대학교 컴퓨터공학과 학사
2009 창원대학교 컴퓨터공학과 석사
2009～현재 창원대학교 컴퓨터공학과 박사과정

관심분야 : 자연어처리, 의견마이닝, 정보검색

차 정 원 (jcha@changwon.ac.kr)

2002 포항공과대학교 공학박사
2003 USC/ISI 박사후 과정
2004～현재 창원대학교 조교수

관심분야 : 기계학습, 자연어처리, 정보검색