Journal of Korea Multimedia Society Vol. 19, No. 8, August 2016(pp. 1564-1573)
http://dx.doi.org/10.9717/kmms.2016.19.8.1564

온라인 활동 데이터를 활용한 영상 콘텐츠의 하이라이트와 검색 인덱스 추출 기법에 대한 연구

하세용[†], 김동환[††], 이준환[†††]

Extraction of Highlights and Search Indexes of Digital Media by Analyzing Online Activity Data

Seyong Ha[†], Dongwhan Kim[††], Joonhwan Lee[†††]

ABSTRACT

With the spread of social media and mobile devices, people spend more time on online than ever before. As more people participate in various online activities, much research has been conducted on how to make use of the time effectively and productively. In this paper, we propose two methods which can be used to extract highlights and make searchable media indexes using online social data. For highlight extraction, we collected the comments from the online baseball broadcasting website. We adopted peak-finding algorithm to analyze the frequency of comments uploaded on the comments section of the website. For each indexes, we collected postings from soap opera forums provided by a popular web service called DCInside. We extracted all the instances when a character's name is mentioned in postings users upload after watching TV, which can be used to create indexes when the character appears on screen for the given episode of the soap opera The evaluation results shows the possibility of the crowdsourcing-based media interaction for both highlight extraction and index building.

Key words: Social Computing, Digital Media Interaction, Highlight, Media Indexing

1. 서 론

인터넷이 널리 보급되고 모바일 기기가 대중화되면서 사람들은 이전보다 더 많은 시간을 온라인 상에서 보내고 있다. 인터넷을 하거나 메시지를 전송하는 일반적인 일에서부터 온라인 쇼핑이나 게임플레이, 소셜 네트워크 서비스(Social Network Service)를 통한 커뮤니케이션 등의 여가 시간 활용에 이르기까지 온라인 활동에 할애하는 시간이 전반적으로 늘어났다. 닐슨 리포트에 따르면 미국인이 일년간 페이스북(Facebook), 블로그(Blog), 텀블러(Tumblr), 트위터(Twitter), 링크드인(LinkedIn)의 다섯 가지 서비스에 소비하는 시간은 모두 약 9억 시간에 달하는데, 이는 파나마 운하나 엠파이어 스테이트 빌딩과 같은 초고층 빌딩을 건설하는데 소요되는 시간보다 훨씬 많은 시간이다. 이처럼 사람들이 온라인에서 소비하는 시간의 양이 방대해지며 이 과정에서 생성된 데이터를 보다 효율적으로 활용하

※ Corresponding Author : Joonhwan Lee, Address: (151-742) Seoul National University, Building 64, IBK Communication Center, 4F, 405, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea, TEL : +82-10-9212-4975, FAX : +82-2-885-8418, E-mail : joonhwan@snu.ac.kr
Receipt date : Jul. 19, 2016, Approval date : Jul. 29, 2016
[†] Dept. of Computer Science, University of Toronto (E-mail : seyongha@cs.toronto.edu)
[††] Dept. of Communication, Seoul National University (E-mail : dongwhan@hcid.snu.ac.kr)
[†††] Dept. of Communication, Seoul National University
※ This study was financially supported in part by the Institute of Communication Research, Seoul National University

고자 하는 방안에 대한 관심이 크게 증가하였다[1].

2006년, 제프 하우(Jeff Howe)는 InnoCentive(이노센티브), 아이스톡포토(iStockPhoto)와 같은 사례를 통해 불특정다수의 개인의 참여가 문제를 해결함에 있어 좋은 결과를 보여준다는 크라우드소싱(Crowdsourcing)의 개념을 세상에 처음 소개했다[2]. 이후, 본 안(von Ahn)이 제시한 휴먼 컴퓨테이션(Human Computation)개념과 같이 컴퓨터 알고리즘의 한계를 집단의 참여를 통해 개선하는 방법[3], 아마존의 메케니컬 터크(Mechanical Turk) 서비스를 이용한 글쓰기 작업에 관한 연구[4], 사용자들이 트위터를 하며 생산해낸 데이터를 바탕으로 미식축구 경기에서 의미 있는 이벤트를 찾는 연구[5] 등 대중들의 직간접적인 참여를 문제 해결에 활용하는 연구들이 시도되었다.

본 연구에서는 이러한 대중들의 참여로 만들어진 활동데이터를 통해 실시간으로 스포츠 경기나 드라마와 같은 영상 콘텐츠의 하이라이트를 추출하거나 주요 등장인물의 검색을 위한 인덱스를 생성하는 방법에 대해 살펴보았다. 영상콘텐츠에서 하이라이트를 추출하거나 특정 인물의 등장 시점을 검색하기 위한 인덱스를 생성하는 일은 관리자가 일일이 영상을 확인해 정보를 입력해야하는 노동집약적인 작업이다. 이러한 문제를 해결하기 위해 이미지 프로세싱 등의 연구가 다수 수행되었지만[6, 7], 영상콘텐츠의 내용을 기반으로 한 알고리즘의 연구는 기술적인 한계로 인해 거의 이루어지지 않았다. 따라서 해당 작업은 여전히 관리자의 시간과 노력에 의존하고 있는 실정이다.

본 연구는 이러한 문제의 해결을 알고리즘의 접근이 아닌 소셜 미디어에서의 사용자 행위 분석을 바탕으로 수행하고자 하였다. 소셜 미디어가 늘어나고 미디어 시청의 장소제한이 사라지며 시청자들은 방송을 시청하면서 동시에 의견을 올릴 수 있는 기회를 갖게 되었고, 경기나 드라마의 흐름에 맞추어 의견을 표출하며 다른 사람들과 커뮤니케이션을 하기 시작했다[8]. 관찰에 의하면 이와 같이 방송을 시청하면서 소셜 미디어에 실시간으로 올리는 글들은 대부분 방송 내용과 밀접한 연관을 갖고 있으며, 이는 방송되는 콘텐츠의 내용을 간접적으로 반영하고 있다고 여겨진다. 본 연구에서는 이러한 시청자들의 실시간 활동 데이터를 수집, 분석해 해당 영상 콘텐츠의 하이라이트를 자동으로 생성하는 방법과 등장인물의 검색 인덱스를 추출하는 방법을 살펴보았다.

2. 관련연구

최근에 온라인 네트워크를 통해 멀리 떨어진 여러 개인들이 서로 의사소통과 협업이 가능하게 됨에 따라 이를 통해 만들어지는 집단적 능력을 활용하여 문제를 해결하고자 하는 시도들이 있는데 이를 집단지성(Collective Intelligence)라 한다. 집단지성이 가장 잘 활용된 예는 위키피디아 백과사전(Wikipedia Encyclopedia)으로, 소수의 교육받은 전문가들에 의해서만 가능하다고 여겨졌던 지식의 축적이 다수의 일반인의 참여로도 가능하다는 것을 보여주었다. 위키피디아는 항목의 숫자 면에서 기존의 백과사전을 압도하는데, 내용의 질이나 신뢰도 차원에서도 브리태니커와 같은 대표적인 백과사전과 비교하여도 크게 떨어지지 않거나 대등하다는 연구 결과도 발표되었다[9]. 이러한 집단지성의 힘은 크라우드소싱, 휴먼 컴퓨테이션, 소셜 컴퓨팅 등의 분야로 확대 되어 갔다. 용어와 뜻에 있어 약간의 차이점이 있긴 하지만 모두 집단지성의 원 의미로부터 파생된 것이다[2].

본 안(von Ahn) 등은 사람들은 컴퓨터가 하지 못하는 일을 쉽게 한다는 점에 착안하여 컴퓨터가 가지는 한계를 집단의 힘으로 극복하고자 한 일련의 연구를 수행하였다. 본 안 등은 '이에스피게임(ESP game)' 연구에서 사람들이 온라인게임 참여를 통해 이미지에 레이블을 붙이는 문제를 해결하는 방법을 제시하였다. 이미지 프로세싱 알고리즘을 활용한 방법은 이미지의 색상, 형태정보 등을 찾아내는 데에는 유용하지만 이미지의 내용을 분석해 구체적인 레이블을 생성하기에는 한계를 가지고 있다. 본 안은 사람들은 이미지가 담고 있는 내용에 기반 하여 레이블을 생성하는데 어려움이 없다는 점에 착안하여 다수의 참여로 이미지 레이블링 작업을 성공적으로 수행해 낼 수 있음을 증명하였다[10]. 또한, 본 안 등은 '리캡챠(reCAPTCHA)' 연구에서 OCR(광학 문자 판독) 알고리즘의 한계를 극복하기 위한 방법을 선보였다[11]. 그동안 OCR 알고리즘은 꾸준한 개선으로 인해 인식율이 상당히 향상되었는데, 그럼에도 불구하고 오래되어 손상되거나 왜곡된 글씨를 읽는 것에는 어려움을 보였다. 일찍이 본 안 등은 웹사이

트 가입 시 봇(bot)을 통한 자동 가입을 방지하기 위해 '캡챠(CAPTCHA)'라는 시스템을 제안하였는데 컴퓨터 알고리즘으로 해독이 불가능한 왜곡된 이미지를 웹사이트 가입 시 사람들에게 보여주고 이를 제대로 입력한 경우에만 가입을 승인하는 방식이었다[12]. 이는 사람들은 글자가 심하게 왜곡 혹은 손상되어도 충분히 읽을 수 있지만 컴퓨터는 그렇지 못하다는 아이디어에서 출발한 것이다. 이 연구는 리캡챠 연구로 발전되어, 책을 스캔할 때 읽지 못하는 문자를 사람들에게 추가로 보여주고 읽게 하여 OCR 알고리즘이 가진 한계를 대중의 힘으로 극복할 수 있다는 사실을 보여주었다[11].

번스타인(Bernstein)은 현재 워드프로세서가 간단하게 제공하고 있는 문장의 교정 및 수정 알고리즘의 한계를 극복하기 위해 일을 작게 나누어 대중에게 맡긴 후 이 과정에서 모은 데이터를 활용하는 크라우드소싱 개념을 적용해 문제를 해결하려는 시도를 선보였다. 이를 위해 아마존의 메케니컬 터그 서비스를 이용해 다수의 많은 참여자를 이끌어 낸 후, 이들의 집합적인 노력을 통해 문장 줄이기나 교정, 수정 등 글쓰기에 도움이 되는 기술을 선보였다[4].

최근에는 본 안, 번스타인의 연구와 같이 대중의 직접적인 참여를 통하여 문제를 해결하는 방법 외에, 트위터 등과 같은 서비스에 사람들이 남기는 활동데이터를 바탕으로 유용한 정보를 만들어내는 연구들이 진행되었다. 마커스(Marcus) 등의 트윗인포(TwitInfo)는 이벤트 시각화 시스템으로 사람들이 트위터에 올린 트윗을 수집하여 분석한 뒤, 현재 일어나고 있는 중요한 이벤트를 찾아준다. 이는 보통 전문가나 특정 매체의 의견을 통해 만들어지는데 트윗인포는 트위터에서 사람들이 올린 메시지를 수집, 분석하여 이를 자동으로 찾아준다[13].

3. 시청자 활동데이터를 활용한 영상콘텐츠 분석

사람들은 종종 TV를 보는 동시에 온라인을 통해 의견을 공유한다. 최근 모바일 기기가 확산되며 TV 시청과 소셜 미디어의 사용이 동시에 이루어지는 경우가 늘고 있는데, 본 연구에서는 영상콘텐츠와 관련해 실시간으로 남겨지는 소셜 데이터를 활용해 하이라이트를 추출하고 주요 등장인물의 검색 인덱스를 자동으로 만드는 방법을 살펴보았다. 본 연구는 알고리즘을 통한 영상콘텐츠의 요소 분석을 넘어 시청자들의 인터넷활동 데이터를 분석하는 소셜 컴퓨팅을 통해 영상콘텐츠의 내용을 분석하고 영상콘텐츠가 가진 의미 있는 부분을 찾아내고자 하는데 그 목적이 있다.

3.1 영상콘텐츠의 하이라이트 생성

국내의 대표적인 포털 서비스인 네이버는 모든 프로야구 경기를 동영상으로 생중계하는 서비스를 제공하고 있다(Fig. 1). 네이버가 제공하는 동영상 중계서비스의 특징 중 하나는 댓글 창이 제공되어 경기를 보는 동시에 응원 글을 올리고, 서로의 의견을 공유할 수 있다는 점이다. 특히, 경기를 실제로 시청하면서 의견을 남길 수 있어 경기와 밀접한 글이 올라오고, 댓글 창이 65자 이내의 단문만을 지원하기 때문에 시청자들이 어떤 장면을 보고 글을 적기까지의 간격이 길지 않아 경기 중에 시청자들의 감정을 움직일 수 있을 만한 장면이 나왔을 때, 시청자들의 댓글이 이러한 점을 즉각 반영하는 모습을 보이곤 한다. 따라서 이들 포털이 제공하는 댓글 창에 시청자들이 남기는 데이터를 이용하여 경기의 하이라이트를 찾아낼 수 있는지를 살펴보았다.

3.1.1 데이터 수집

본 연구에서는 한국프로야구 정규경기 중 3경기를 선택, 각 경기의 네이버 스포츠 중계센터의 댓글 창에 올라온 사용자 댓글 데이터를 수집하였다. 네이버의 중계서비스는 HTML이라는 웹문서형식으로 표현되어진다. 특히, Fig. 1에 보이는 것처럼 네이

Fig. 1. Naver Sports Center Screen. Viewer's comments are represented in a red box.

버 스포츠 중계센터 댓글창은 일정한 문서형식을 가지고 있다. 댓글창은 댓글의 리스트형식으로 이루어져있고, 각 댓글은 댓글내용, 비공개된 작성자아이디, 작성시간, 그리고 응원팀이라는 동일한 형식으로 나타내진다. 연구에서는 HTML문서의 수집과 그 문서분석을 위해 루비(Ruby) 언어로 크롤링 소프트웨어를 작성하였다. 크롤링 소프트웨어는 댓글들이 나타나는 댓글창의 HTML문서들을 수집한 뒤, HTML문서의 구문분석을 통해, 댓글창에 올라온 댓글 중 댓글내용, 작성시간, 응원팀만을 추출하여 데이터를 저장하였다. 프로야구경기는 시작시간은 정해져있으나, 종료시간이 경기마다 상이하기에 데이터 수집시 경기 시작 시간부터 그날 자정까지의 데이터만으로 한정하였다.

3.1.2 최고점 찾기 알고리즘과 하이라이트 추출

댓글 데이터는 히스토그램(Histogram)과 마커스 등이 트윗인포(TwitInfo) 연구에서 제시한 최고점 찾기 알고리즘을 사용하여 분석하였다[13]. 여기서 사용된 히스토그램 분석은 매 분마다 계산한 댓글의 빈도수 데이터를 최고점찾기 알고리즘을 이용해 추출하는 방법으로 영상콘텐츠의 하이라이트 구간을 찾는데 사용되었다(Fig. 2).

시청자들의 감정의 변이 추이가 하이라이트와 밀접한 연관이 있을 것으로 가정하였기 때문에, 히스토그램 상 최고점(peak)인 구간, 즉, 댓글이 가장 많이 달렸던 시간대가 시청자들이 느끼는 경기의 하이라이트 구간을 반영할 것이라고 가정하였다. 그러나 온라인 중계방송의 특성상 시청자의 유입이 고정되어있지 않기 때문에, 전체데이터를 기준으로 최고점을 찾는 것은 경기의 전반부와 같이 사용자의 참여가 상대적으로 부족한 구간에서의 하이라이트를 반영하기에 불충분할 수 있다. 예를 들어, 경기초반보다 경기후반에 댓글 수가 2배 이상 증가한다면, 경기 전체시간을 기준으로 최고점을 찾아 하이라이트로 가정한다면 전반부의 하이라이트 장면을 놓칠 가능성이 높다. 따라서 일정시간 범위 안에서 상대적으로 댓글이 증가한 구간을 찾아야 추출된 하이라이트의 정확성을 높일 수 있다. 본 연구에서 사용한 트윗인포의 최고점찾기 알고리즘은 이러한 점을 고려한 알고리즘으로 연속적인 시계열 상의 데이터분포에서 최고점 구간을 찾는 알고리즘이다[13]. 단순히 시계열 상의 데이터분포에서 가장 높은 점 하나를 찾는 것이 아니라, 어떤 한 시점에서의 데이터 값과 그 이전까지 데이터 값을 고려하여 상대적으로 최고점인 구간들을 찾는다. 하지만, 이렇게 찾아낸 최고점 구간은 전체 영상에서의 한 순간을 의미하기에 본 연구에서는 최고점 구간을 중심으로 전후 1 분씩을 추가하여 하이라이트 구간으로 설정하였다.

3.1.3 하이라이트 추출 결과

프로야구경기 중 선택된 3경기의 경기 당 평균 댓글의 수는 10881.3개 (±7493.11) 이었다. 댓글 수의 편차가 큰 이유는 팀 별 인기도에 따른 팬 수, 경기가 열린 시기와 장소에 따른 관중의 수, 그리고 그날 경기의 볼거리 여부에 따른 차이로 해석할 수 있다. 즉, 득점없이 단조로운 흐름의 경기이거나 비인기구단의 경기는 상대적으로 댓글이 적을 거라 예상할 수 있다.

추출된 하이라이트 장면은 경기 당 평균 13개

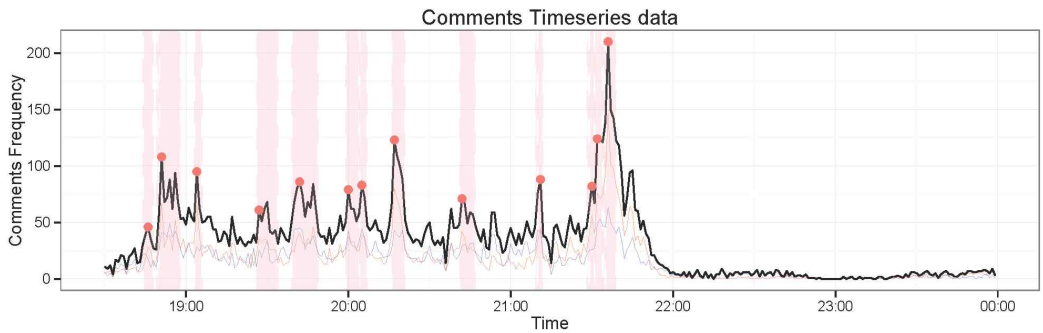

Fig. 2. Comments Frequency - Time graph. Each line represents comments frequency of total(home+away), home, away team. Peaks from peak-finding algorithms depicted as dots.

$$\mathrm{Precision} = \frac{|relevant-documents| \bigcap |retrieved-documents|}{|retrieved-documents|}$$

Fig. 3. Precision Formula.

(±2.65)로 네이버 스포츠 중계센터가 제공하는 하이라이트 숫자의 평균인 37개보다 절반 이상 적었다. 추출된 하이라이트 장면을 분석해보면, 득점과 관련된 장면이 대부분을 차지했으나, 모든 득점 장면이 추출되지는 않았고, 득점 장면들 중 동점 장면이나 역전 장면 등 상대적으로 중요한 장면들 위주로 추출되었음을 확인할 수 있었다. 득점 장면 이외에는 경기의 흐름에 영향을 준 실책 혹은 호수비 장면들과 같이 시청자들이 같이 크게 반응할만한 장면들이 추출되었다.

3.1.4 하이라이트의 평가

하이라이트란 스포츠에서 가장 중요하거나 흥미있는 장면을 뜻한다. 야구 경기에서는 응원팀의 득점장면이 하이라이트가 될 수도 있지만, 개인이 경기를 관람하는 관점에 따라, 실책이나 호수비등, 개개인마다 하이라이트라고 생각하는 장면이 상이할 수 있다. 그렇기 때문에, 본 연구에서는 정보검색에서 사용하는 '정확률' 의 개념을 가져와 하이라이트 추출방법을 평가하였다. 정확률이란 정보검색에서 사용되는 평가방법 중 하나로, 사용자가 입력한 검색어에 대해 검색엔진의 검색결과가 사용자가 찾고자하는 결과와 관련 있는 문서를 얼마나 포함하고 있는지를 나타낸다(Fig. 3) 이를 본 연구의 평가방법으로 적용하면, 먼저, 네이버 스포츠 중계센터에 올라와있는 하이라이트 장면들을 일반적으로 사람들이 접하는 야구경기 하이라이트라 가정한다(relevant_documents). 많은 사람들, 특히 중계센터를 통해 스포츠를 관람하는 사람들은, 네이버 스포츠 중계센터에 올라와있는 하이라이트 장면을 감상하기 때문에 그들이 일반적으로 접하는 하이라이트가 네이버 중계센터에서 제공하는 하이라이트라 가정할 수 있다. 만약, 연구에서 제시한 하이라이트 추출방법으로 추출한 하이라이트 장면들(retrieved_documents)이 네이버 스포츠 중계센터에서 추출한 하이라이트 장면을 많이 포함하고 있다면, 연구에서 제시한 추출방법이 사람들이 일반적으로 공감할 만한 장면을 하이라이트로 뽑았다고 이야기할 수 있다.

네이버 스포츠 중계센터에서 제공하는 하이라이트와의 비교 분석 결과는 다음과 같다. 네이버 스포츠 중계센터에서 제공하는 하이라이트는 평균 6~7분 정도로, 해당 경기의 득점 장면을 위주로 제공하고 있다. 경기 당 제공한 하이라이트는 평균 37.3개의 장면으로 장면 당 평균 15초 정도를 할애하고 있다. 제공되는 하이라이트 구간의 길이는 장면에 따라 다양했지만 평균적으로 본 연구에서 추출한 하이라이트의 길이가 더 길었기 때문에 장면 비교 시, 추출한 하이라이트 구간이 네이버의 하이라이트 장면을 포함하고 있으면 같은 하이라이트를 추출했다고 가정하였다.

Table 1은 각 하이라이트 추출에 관한 결과 값이다. 이를 바탕으로 한 계산한 정확률의 평균은 약 52%로 본 연구에서 추출한 하이라이트의 절반 정도가 네이버에서 선정한 하이라이트 장면을 포함하고 있었다. 이는 네이버에서 제공하는 하이라이트가 거의 대부분 득점 장면 위주로 제공되는데 비해 본 연구에서 추출한 하이라이트는 시청자들이 중요하다고 여기는 득점 장면 들 위주로 추출되었고, 그 밖에 네이버에서는 제공하지 않는 팬들이 선호하는 경기 장면 (예: 호수비, 벤치클리어링 등) 이 포함되었기 때문으로 해석할 수 있다. 각 하이라이트 별로 사람들의 반응정도가 다를 수 있기에, 추출된 하이라이트 구간의 길이 대비 댓글이 빨리 증가한 상위 50%의 하이라이트를 따로 뽑아 비교 분석한 결과 정확

Table 1. Analysis result of extracted highlight scenes

	Number of comments	Number of extracted highlight scenes	Precision	Top 50% precision
game 1	18961	15	53.3%	87.5%
game 2	9523	15	53.3%	75%
game 3	4160	10	50%	80%

률은 약 80%로 전체 하이라이트와 비교해서 다소 높게 나타났다. 이는 사람들이 선호하는 하이라이트의 장면이 네이버 등이 제공하는 득점 장면에만 머물러 있지 않음을 확인하게 해준다.

3.2 동영상 검색 인덱스 생성

본 연구의 두 번째 단계에서는 소셜 데이터를 통해 영상콘텐츠의 검색 인덱스를 자동으로 생성하는 방법에 대한 연구를 수행하였다. 이 연구 역시, TV 시청 중 온라인에 남기는 글을 분석하여 수행되었다. 최근 모바일 기기의 사용이 급격하게 증가하면서 TV드라마 등을 시청하며 시청의견 등을 소셜 미디어를 통해 다른 사람과 공유하는 경우가 종종 목격되곤 한다. 외국의 경우, TV 프로그램에 대한 시청자들의 의견공유가 주로 트위터를 통해 활발히 일어나고 있는데 해시태그(hash-tag)라는 대중분류법(folksomony) 장치를 사용하여 특정 TV 프로그램과 관련한 이야기를 공유한다. 국내의 경우에 이러한 의견공유는 트위터와 같은 소셜 미디어보다는 인터넷 커뮤니티 게시판을 통해 활발히 이루어지고 있다. 국내의 여러 인터넷 커뮤니티 게시판 중 디시인사이드 (DC Inside) 커뮤니티는 다른 커뮤니티와는 다르게 특정 드라마에 대해 토론할 수 있는 게시판을 '갤러리'라는 이름으로 제공한다. 모든 드라마에 대해 갤러리가 생기지는 않지만, 특별히 화제가 되었던 드라마를 중심으로 만들어지는 갤러리를 살펴볼 수 있었다. 이 게시판의 주 이용자들은 드라마 시청자들이며, 보통 방영시간 대와 방영이 막 끝난 시점에 집중적으로 글을 올리는 특징을 보인다. 관찰한 바에 의하면 방영 중 실시간으로 올라오는 글은 현재 방영되는 드라마의 특정 장면과 관련한 내용이 대부분이었다. 따라서 본 연구에서는 이와 같이 드라마의 방영시간 중에 사람들이 올린 글들의 내용을 분석하여 검색을 위한 인덱스로 활용할 수 있는지를 살펴보았다.

3.2.1 데이터 수집 및 인물사전 작성

데이터의 수집을 위해 디시인사이드 갤러리 게시판 중 공중파 드라마 '골든타임' 중 7회분을 선정하여 분석하였다. 하이라이트 추출의 경우와 마찬가지로 루비 언어로 제작된 크롤링 소프트웨어를 사용하여 데이터를 수집하였다. 데이터의 수집은 방영 후 12시간이 지나고 나서 작성자, 작성시간, 작성 글을 중심으로 이루어졌다. 게시판 글은 드라마 공식 시작 시간부터 종료하는 시간 사이에 작성된 글로 한정하였다.

본 연구에서는 특정 배우가 등장하는 장면을 검색할 수 있는 인덱스를 자동으로 추출하는 것을 목표로 하였다. 이에 데이터의 수집 및 분석은 주로 등장 배우가 언급되는 순간을 추출하는 방식으로 이루어졌다. 그러나 게시판에 올라오는 글들의 특징을 살펴보면 배우들의 언급이 매우 다양하게 이루어지고 있음을 알 수 있다. 사람들의 게시글 속에서 배우들은 극 중 인물의 이름뿐만 아니라, 배우의 본명, 시청자들의 애칭 등 여러 이름으로 불린다. 또한, 표준 맞춤법을 사용하지 않은, 이른바 통신언어나 속어 등의 표현도 다수 발견된다. 예를 들어, 의학드라마 '골든타임'의 주인공인 '박민우(본명 이선균)' 의 경우, 게시판 글에서는 극 중 이름인 박민우 외 에도 '이선균', '민우쌤' 등으로 불린다. 이처럼 하나의 인물이 다양한 표현으로 지칭되기 때문에 검색 인덱스의 정확도를 높이기 위해서는 인물을 지칭하는 모든 표현을 아우르는 참고자료가 필요하다.

이에 본 연구에서는 우선 등장인물에 대한 '인물사전'을 작성하였다. 디시인사이드의 드라마 게시판에는 '단어장'이라는 이름으로 제공되는 인물들의 별칭 목록이 있다. 이 별칭 목록은 이제껏 팬들이 게시판에서 사용한 별칭들을 모아 놓은 것으로 게시판 글에서 언급되는 인물들의 이름을 나열하고 있다. Fig. 4에서 나타나듯이 본 연구에서 제작된 인물사전은 디시인사이드의 단어장을 바탕으로 드라마 속 인물의 이름과 배우의 실제 이름, 게시판에서 인물을 언급할 때 사용되는 별칭 등으로 이루어진 텍스트파일의 형태로 만들어졌다.

3.2.2 검색 인덱스 생성

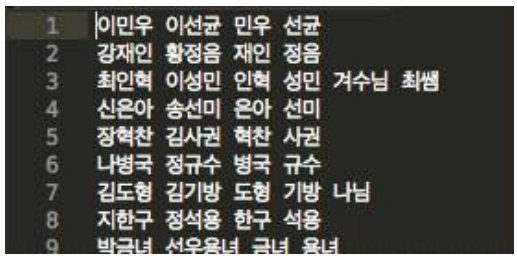

Fig. 4. Character Thesaurus.

검색 인덱스의 생성은 수집된 게시글 속에서 특정 인물이 언급되었는지의 여부를 확인하여 수행되었다. 먼저 작성된 인물사전을 기초로 하여, 수집된 게시 글 데이터에서 해당 인물이 언급된 게시 글을 검색하였다. 게시판 데이터로부터 인물의 이름이 검색되면, 해당 인물의 실제이름과 인물사전에 등록된 이름들, 그리고 게시판 글의 원본과 작성된 시간 등을 저장하였다. 인물사전에 등록된 이름들과 해당 게시 글의 원본은 차후에 검색결과를 검증하기 위한 수단으로 사용되었다. 검색 인덱스로 사용되는 데이터는 검색된 이름과 게시 글이 올라온 시간이다. 사용자가 게시 글에서 어떤 인물에 대한 언급을 했을 경우, 그 시간대에 드라마에서 해당 인물이 나왔다고 가정하고 진행하였다.

3.2.3 검색 인덱스 생성 및 검증

드라마가 방영되는 중에 게시판에 올라온 글은 평균 265개로 분당 평균 4.03개였다. 게시판에 올라온 글 중 인물사전에 등록된 이름을 내포하고 있는 글은 평균 90개 정도로 조사되었는데, 이는 전체 게시글 중 34.19% 정도에 해당되는 글이다. 이를 이용하여 검색 인덱스를 생성하였다.

Table 2는 에피소드 별로 수집된 게시물의 데이터이다. 표를 보면 에피소드 별로 수집된 게시글의 편차가 큰 것을 볼 수 있다. “방송 중 게시된 글의 수”와 “인덱스로 사용된 게시글의 수” 와의 상관관계는 0.95로, 게시글이 많이 올라올 수록 인덱스로 사용되는 글이 많아짐을 나타낸다.

하이라이트의 검증과 동일한 방법으로 수행한 인덱스의 정확률 검증 결과는 Table 2에서 볼 수 있듯이 70%를 상회하는 높은 결과를 보여주었다. 인덱싱의 정확율은 정확도 분석 구간과 관련이 있었다. 본 연구에서는 정확도 분석 구간을 30초로 가정하고 연구를 진행하였는데, 이는 사람들이 특정 장면을 본 후 해당 장면에 대한 글을 작성하고 올리기까지의 시간을 고려한 것이다. 사용자의 다양한 컴퓨팅 환경을 고려한다면 30초라는 값은 정확하다고 볼 수는 없지만, 이 구간에서도 충분히 신뢰도 있는 정확율을 보여주었다. 또한, 특정 배우가 등장하는 하나의 장면의 단위가 보통 최소 30초 정도로 조사되었기에 이를 기본 단위로 하여 분석이 진행되었다.

정확도 분석 구간의 길이를 40초, 60초로 늘여 분석을 해 보았을 때 더 높은 인덱싱의 정확율이 발견되었다. 예를 들어 에피소드 7의 경우, 정확도 분석 구간을 60초로 지정하여 분석한 결과 인덱스의 정확율이 19% 가량 증가하였다. 방송 시청 중 글을 포스팅 할 수 있는 효율적인 인터페이스를 제공할 수 있다면 정확도 분석 구간의 길이를 보다 정확하게 예측할 수 있을 것이나 본 연구에서는 사용자의 다양한 행위가 발생하는 시간 차이의 변인을 통제하지 않았기에 시각 자극을 받아들인 후에 이루어지는 행위 (예: 글의 게시) 까지 걸리는 시간을 연구대상에 포함시키지는 않았다.

4. 논의 및 향후 연구 방향

본 연구에서는 소셜 컴퓨팅을 통한 영상콘텐츠의 분석 방법에 대해 살펴보았다. 먼저 TV시청자들의 온라인 활동 데이터를 이용하여 자동으로 하이라이트 장면을 찾아내고, 게시판에 실시간으로 올린 게시글의 내용을 분석하여 등장인물의 출연 위치를 검색할 수 있는 인덱스를 만드는 방법을 제안하였다. 이를 위해 웹크롤링 소프트웨어를 제작하여 야구와 드라마에 대한 시청자들의 온라인 반응 데이터를 수집한 후, 히스토그램 분석과 최고점 찾기 알고리즘

Table 2. Result of building search indexes

	Number of total posts	Number of posts mentioning characters	Number of posts per min	Precision
Episode 1	226	90	3.23	72%
Episode 2	306	100	4.37	83%
Episode 3	384	114	5.49	77%
Episode 4	260	75	3.71	64%
Episode 5	441	164	6.30	59%
Episode 6	87	25	1.24	80%

을 적용하여 스포츠 경기의 하이라이트를 자동으로 추출하였고, 미리 제작한 인물사전을 이용한 검색을 통해 드라마 등장인물의 검색 인덱스를 자동으로 추출하였다.

정확율 검증 결과는 각각 하이라이트 자동 추출이 네이버의 상위 50% 하이라이트 대비 80% 이상, 검색 인덱스 생성은 72% 이상의 결과를 보여주었다. 비록, 이 두 가지 기법이 기존의 동영상 콘텐츠의 하이라이트 추출과 검색 인덱스 생성을 완전히 대체할 수는 없을지라도, 관리자가 수작업으로 수행하던 작업을 보완하는 도구로 충분한 효율을 가지고 있음을 보였다.

본 연구의 또 다른 시사점은 비자발적으로 생성된 데이터를 활용하여 기존의 기술이 가진 한계를 극복하고자 시도한 바에 있다. 본 연구에서는 시청자들의 온라인 상 에서의 일상적인 활동의 분석을 통해 중요한 정보가 무엇인지를 찾아낼 수 있음을 밝히고 있다. 이 과정에서 본 연구는 사람들에게 의도적으로 게시 글을 남겨달라는 요청이나 연구에 참여를 부탁하지 않았다. 예를 들어, 온라인으로 설문조사를 실행하고 참여하는 서비스인 아마존의 메케니컬 터크는 자발적인 참여의사를 밝히는 다량의 사용자가 있을 때에만 원활한 조사가 이루어진다. 하지만 본 연구에서는 사용자들이 TV를 보며 일상적으로 남기는 행적을 추적하여 이를 중요한 데이터로 활용했다는 점에서 의의가 있다.

본 연구를 진행하면서 몇 가지의 미비한 점이 발견되었고, 향후 이에 대한 보완 연구가 진행될 예정이다. 먼저, 하이라이트의 추출 결과는 네이버 하이라이트와 비교해 정확율의 차이가 많지 않았으나, 실제로 추출된 하이라이트의 만족도와 신뢰도는 조사하지 못했다. 이를 위해 추가 연구를 진행할 계획이다. 또한 미디어 인덱싱 과정에서도 형태소 분석 등을 통해 보다 정교한 키워드의 추출 작업이 필요할 것이다. 본 연구를 진행하면서 가장 제약점으로 다가왔던 부분은 데이터의 크기이다. 스포츠 경기 하이라이트의 경우 어느 정도 수용 가능한 크기의 데이터가 수집되었지만, 드라마 검색 인덱스 연구의 경우 실시간으로 의견을 교환하는 활성화된 인터넷 커뮤니티가 그리 많지 않아 상대적으로 적은 수의 데이터를 수집하는데 그쳤다. 적은 데이터로도 약 72%의 정확율을 보여주었다는 사실은 더 많은 데이터가 수집될 경우, 정확율도 더 높아질 수 있다는 가능성을 보인 것이라 생각된다. 향후, 트위터 등을 통한 의견교류가 훨씬 활발한 외국의 드라마나 스포츠 중계 등의 분석을 통해 본 연구가 보여준 가능성을 확인해 볼 예정이다.

본 연구의 결과는 다른 관련 연구로의 확장도 가능해 보인다. 스포츠 경기와 같이 시청자 층이 확실하게 나뉘는 경우에는 팀별로 열광하는 장면과 화제가 되는 장면이 다르다는 가정 아래 각 팀별 하이라이트를 추출하는 것도 가능하다. 미디어 인덱싱의 경우에도 현재의 연구는 특정 인물이 등장하는 시점의 인덱스를 찾아내는데 그치고 있으나, 개별 등장인물이 언급되는 빈도수를 따로 계산한다면 해당 인물이 얼마나 자주 등장하는지의 여부와 상관없이 시청자들에게 얼마나 회자되고 있는지를 파악할 수 있어 각 등장인물의 극중 가시성(visibility)을 계산해 낼 수 있다.

5. 결 론

시각정보와 관련한 선행 연구에 따르면 시각적으로 전달된 정보는 노출된 지 1초 후에 급속하게 기억에서 사라지게 된다[14-16]. 그렇기 때문에 사람들이 TV를 시청하면서 게시판에 글을 남기는 행위는 무작위적 행위라고 볼 수 없다. 시각 자극이 도달한 후 게시판에 글을 올리기까지 어느 정도 시간이 걸린다는 것을 감안한다면 사람들은 TV시청 중에 본 여러 시각 정보들 중에 기억에서 사라지지 않고 남아있는 정보를 게시판에 남기게 된다고 여겨진다. 따라서 이들 정보는 여러 사라진 시각 정보들과 비교하여 의미를 가지며 만일 불특정 다수가 동일한 정보를 남기게 된다면 이들 정보의 중요성은 더욱 커질 수 있다. 본 연구는 소셜 컴퓨팅을 통해 이러한 정보를 영상콘텐츠 분석에 활용하고자 하였다.

이를 위해 TV시청자들의 온라인 활동 데이터를 수집하여 하이라이트 구간을 뽑아내고, 영상콘텐츠의 검색 인덱스를 만드는 새로운 방법에 대해 소개했다. 또한 본 연구에서는 결과의 신뢰성을 측정하기 위해 정확율 공식을 자동으로 추출된 하이라이트와 검색 인덱스를 검증하였다. 그 결과, 시청자들의 온라인 활동들을 모아 신뢰할 수 있고 만족스러운 하이라이트와 검색 인덱스를 만들어낼 수 있음이 확인되었다.

본 연구의 결과는 현재 많이 논의가 진행되고 있는 인터랙티브 TV 또는 스마트 TV의 서비스 개선에도 매우 중요한 시사점을 제공하고 있다. 앞으로 도입될 스마트 TV는 인터넷과 연결되어 다양한 온라인 서비스를 사용할 수 있도록 지원할 것으로 예상된다[17]. 서비스 설계 시에 이러한 사용자 행위 데이터를 활용할 다양한 방법을 모색한다면, 그동안 기술적으로 어렵게만 여겨졌던 동영상의 검색 키워드, 하이라이트 등을 자동으로 추출할 수 있는 중요한 데이터들을 확보할 수 있을 것이다.

REFERENCE

[1] NIELSEN: State of the Media: The Social Media Report Q3 2011, http://www.nielsen.com/us/en/reports/2011/social-media-report-q3.html, (accessed Nov., 30, 2012).

[2] J. Howe, *The Rise of Crowdsourcing*, *Wired Magazine*, Vol. 14, No. 6, pp. 1-4, 2006.

[3] L.V. Ahn, "Human Computation," *Proceeding of Design Automation Conference*, pp. 418-419, 2009.

[4] M.S. Bernstein, G. Little, R.C. Miller, B. Hartmann, M.S. Ackerman, D.R. Karger, et al., "Soylent: A Word Processor with a Crowd Inside," *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology,* pp. 313-322, 2010.

[5] A. Tang and S. Boring, "#EpicPlay: Crowdsourcing Sports Video Highlights," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1569-1572, 2012.

[6] P. England, R.B. Allen, and M. Sullivan, "I/Browse: the Bellcore Video Library Tool Kit," *Proceeding of Electronic Imaing: Science & Technology*, pp. 254-264, 1996.

[7] M. Gelgon and P. Bouthemy, *Determining a Structured Spatio-temporal Representation of Video Content for Efficient Visualization and Indexing*, Springer, Berlin Heidelberg, 1998.

[8] K. McPherson, K. Huotari, F. Cheng, D. Humphery, C. Cheshire, and A.L. Brooks, "Glitter: A Mixed-Methods Study of Twitter Use During Glee Broadcasts," *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion,* pp. 167-170, 2012.

[9] J. Giles, "Internet Encyclopaedias Go Head to Head," *Journal of Nature,* Vol. 438, No. 7070, pp. 900-901, 2005.

[10] L.V. Ahn and L. Dabbish, "Labeling Images with a Computer Game," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319-326, 2004.

[11] L.V. Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum, "Recaptcha: Human-based Character Recognition via Web Security Measures," *Journal of Science,* Vol. 321, No. 5859, pp. 1465-1468, 2008.

[12] L.V. Ahn, M. Blum, N. Hopper, and J. Langford, "CAPTCHA: Using Hard AI Problems for Security," *Proceeding of Advances in Cryptology-EUROCRYPT,* pp. 646-646, 2003.

[13] A. Marcus, M.S. Bernstein, O. Badar, D.R. Karger, S. Madden, and R.C. Miller, "Twitinfo: Aggregating and Visualizing Microblogs for Event Exploration," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 227-236, 2011.

[14] G.W. Allport, "Change and Decay in the Visual Memory Image," *British Journal of Psychology General Section*, Vol. 12, Issue 2, pp. 133-148, 1930.

[15] E. Averbach and A.S. Coriell, "Short-term Memory in Vision," *Bell System Technology Journal,* Vol. 40, Issue 1, pp. 309-328, 1961.

[16] D.E. Broadbent, *Perception and Communication*, *Pergamon Press*, London, 1958.

[17] J. Lee, C. Rim, W. Kim. "A Study of Design and Implementation of Software Player for XML Data Based Interactive Digital TV Service," *Journal of Korea Multimedia Society*, Vol. 7, Issue 11, pp. 1564-1570, 2004.

하 세 용

2011년 2월 한양대학교 에리카캠퍼스 컴퓨터공학과 학사
2015년 2월 서울대학교 협동과정 인지과학전공 석사
2015년~현재 University of Toronto, Dept. of Computer Science 박사과정
관심분야 : End-user programming, Pen-computing, Social Computing

김 동 환

2006년 5월 Long Island University Computer Science 학사
2007년 8월 카네기멜론 대학교 Human-Computer Interaction 석사
2009년 4월 Wireless Generation(New York City, USA), Product Designer
2012년 7월 LG전자 MC연구소 UX 디자이너
2012년 9월~현재 서울대학교 언론정보학과 박사과정
관심분야 : HCI, Computational Journalism, Social Computing

이 준 환

1995년 2월 서울대학교 산업디자인학과 학사
2000년 5월 카네기멜론 대학교 인터랙션 디자인 전공 석사
2008년 5월 카네기멜론 대학교 School of Computer Science 박사
Human-Computer Interaction 전공
2011년~현재 서울대학교 언론정보학과 부교수 (HCI+D Lab.)
관심분야 : HCI, Social Computing, Information Visualization, Interaction Design