

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

논문 출처: <https://dbpia.co.kr/journal/articleDetail?nodeId=NODE07002180>

'네이버 야구 중계'에서 실시간 댓글 창에 올라오는 댓글의 개수를 이용해서 하이라이트를 추출하고, 미리 인물 사진을 만들고 이를 이용해서 인터넷 커뮤니티에 올라오는 게시물을 검색해서 검색 인덱스를 추출했음.

<하이라이트 추출>

1. 데이터 수집

국내의 대표적인 포털 서비스인 네이버는 프로야구 생중계 서비스는 실시간으로 댓글을 달 수 있다. 이를 크롤링하여 실시간 댓글 데이터를 수집함. (롤의 경우는 네이버에서 실시간 댓글 서비스를 제공하지 않으므로 트위치 등을 이용해서 할 예정)

댓글창에 올라 온 댓글 중 댓글내용, 작성시간, 응원팀을 추출하여 데이터를 저장하였다.

2. 최고점 찾기 알고리즘과 하이라이트 추출

댓글이 가장 많이 달린 시간대가 시청자들이 느끼는 하이라이트 구간을 반영할 것이라고 가정.

매 분마다 댓글의 빈도수를 계산.

하지만, 전체데이터를 기준으로 최고점 을 찾는 것은 경기의 전반부와 같이 사용자의 참여가 상대적으로 부족한 구간에서의 하이라이트를 반영하기에 불충분할 수 있다.

- ⇒ 일정시간 범위 안에서 상대적으로 댓글이 증가한 구간을 찾음.
- ⇒ 히스토그램과 마커스등이 트윗인포(TwitInfo) 연구에서 제시한 최고점 찾기 알고리즘 (연속적인 시계열 상의 데이터분포에 최고점 구간을 찾는 알고리즘이다)을 사용.

시계열 상의 데이터분포에서 어떤 한 시점에서의 데이터 값과 그 이전까지 데이터 값을 고려하여 상대적으로 최고점인 구간들을 찾는다.

최고점 구간을 중심으로 전후 1 분씩을 추가하여 하이라이트 구간으로 설정.

3. 하이라이트 추출 결과

- 득점없이 단조로운 흐름의 경기이거나 비인기구단의 경기는 상대적으로 댓글이 적을 거라 예상

할 수 있다.

- 추출된 하이라이트 장면을 분석해보면, 득점과 관련된 장면이 대부분 추출 되었고, 득점 장면들 중에서도 동점 장면이나 역전 장면 등 상대적으로 중요한 장면들 위주로 추출됨. 득점 장면 이외에는 경기의 흐름에 영향을 준 실책 혹은 호수비 장면들과 같이 시청자들이 같이 크게 반응할 만한 장면들이 추출됨

4. 하이라이트의 평가

네이버 스포츠 중계센터에서 추출한 하이라이트 장면이 연구에서 추출한 하이라이트 장면과 유사한지를 평가함(연구에서 추출한 최고점이 네이버 하이라이트 장면에 얼마나 포함되어 있는지)

정확률:

전체 하이라이트 영상 대비 평균은 약 52%.

네이버 생중계에서 실시간 댓글 수가 폭발적으로 증가한 상위 50% 하이라이트 대비 80% 이상.

결과 분석: 장면 들 위주로 추출되었고, 그 밖에 네이버에서는 제공하지 않는 팬들이 선호하는 경기 장면 (예: 호수비, 벤치클리어링 등) 이 포함되었기 때문으로 해석

<동영상 검색 인덱스 생성>

1. 목표

소셜 데이터를 통해 영상콘텐츠의 검색 인덱스를 자동으로 생성하는 방법에 대한 연구

목표: 드라마에서 특정 배우가 등장하는 장면을 검색할 수 있는 인덱스를 자동으로 추출.

드라마의 방영시간 중에 사람들이 인터넷 커뮤니티에 올린 글들의 내용을 분석하여 검색을 위한 인덱스로 활용할 수 있는지를 살펴봄

2. 데이터 수집

사용자가 게시 글에서 어떤 인물에 대한 언급을 했을 경우, 그 시간대에 드라마에서 해당 인물이 나왔다고 가정하고 진행

- 드라마 '골든타임' 중 7회분을 선정하여 분석
- 12시간이 지나고 나서까지 작성 글을 중심으로 수집
- 등장 배우가 언급되는 시간을 추출

=> 하나의 인물이 다양한 표현으로 지칭되기 때문에 검색 인덱스

의 정확도를 높이기 위해서는 인물을 지칭하는 모든 표현을 아우르는 참고자료가 필요

=> 등장인물에 대한 '인물사전'을 작성

1	이민우	이선균	민우	선균
2	강재인	황정음	재인	정음
3	최인혁	이성민	인혁	성민
4	신은아	송선미	은아	선미
5	장혁찬	김사권	혁찬	사권
6	나병국	정규수	병국	규수
7	김도형	김기방	도형	기방
8	지한구	정석용	한구	석용
9	박국년	선우용년	국년	용년

- 인물사전을 기초로 하여, 수집된 게시 글 데이터에서 해당 인물이 언급된 게시 글 검색.

- 게시글이 작성된 시간의 30초 이내가 그 등장인물이 등장한 시점이라고 가정하고 인덱스 생성. (시간은 40초, 60초로 늘려가며 실행함)

정확도는 70%.

!! 트위치 등에서 제공되는 룰의 실시간 댓글창을 이용해서 이미지 분석을 통해 추출한 하이라이트를 보조할 수 있다고 판단.

!! 검색 인덱스 같은 경우는 이 논문에서는 커뮤니티를 이용했지만 실시간 댓글창에 언급되는 인물 / 포지션등을 이용해서 추출할 수 있다고 생각.