웹 페이지 구조 분석을 통한 효과적인 동영상 검색용 키워드 추출 방법

이종원○ 최기석 장주연 낭종호
청강문화산업대학 e스포츠게임과○, 서강대학교 컴퓨터학과
jongwon○@ck.ac.kr, {gschoi, jyjang, jhnang}@sogang.ac.kr

An Effective Keyword Extraction Method Based on Web Page Structure Analysis for Video Retrieval in WWW

Jongwon Lee○ Giseok Choi Juyeon Jang Jongho Nang
Chungkang College○ Sogang University

웹 기반의 멀티미디어 데이터 중에서 웹 이미지 관리 및 검색 시스템에 대해서는 활발한 연구가 진행되어 왔으나, 동영상은 파일 크기가 매우 크고, 웹에 많이 분포해 있지 않았을 뿐만 아니라, 몇 개의 특정 동영상 제공 업체에서 보유하는 동영상들이 대부분을 차지하여왔기 때문에 관련 연구가 활발하게 진행되지 않았다. 그러나 최근 들어 이전보다 더 많은 웹 사이트에서 동영상 컨텐츠를 제공하고 있고, UCC처럼 일반 사용자들도 쉽게 동영상을 웹 페이지에 게시할 수 있게 됨에 따라 웹 기반 동영상의 효율적인 검색기능이 요구 되고 있다.

동영상과 이미지는 단순히 파일 포맷에서의 차이뿐만 아니라, 웹 페이지 내에서 미디어 데이터가 차지하는 의미가 다르기 때문에 같은 주석 방법을 사용하는 것은 적합하지 않다. 이미지는 웹 페이지 내에서 의미적으로 중심이 되는 것이 아니라 텍스트를 위한 보조 수단의 역할을 하는데 비해, 동영상은 웹 페이지 내에서 중심이 되고, 주변 텍스트들이 동영상 데이터를 기술하기 위한 보조 수단이 된다. 이와 같이 웹 동영상 페이지는 웹 이미지 페이지와 다른 특성을 가지고 있으므로, 키워드 추출에 있어서 기존과는 다른 방법이 요구된다.

웹 페이지로부터 보다 정확한 키워드를 추출하기 위해서는 웹 페이지 내에서 동영상과 상대적으로 상관관계가 높은 텍스트 영역을 찾고, 그 영역으로부터 키워드를 추출해야 한다. 따라서 동영상과의 상관관계가 높은 텍스트 영역을 찾으려면, 먼저 웹 페이지에 있는 텍스트 영역을 분리하여 동영상과의 레이아웃 거리(Layout distance)를 측정한다. 레이아웃 거리를 계산하기 위해서 VIPS DOM 트리를 이용한다. 웹 페이지의 VIPS DOM 트리에서 동영상노드로부터 폭 방향으로 멀리 떨어져 있을수록 화면상의 거리가 멀어지므로 레이아웃 거리는 커지게 된다. 그러나 DOM트리에서 깊이가 깊어지면 화면상의 배치가 세밀하고 복잡해질 뿐, 거리가 멀어지는 것이 아니므로 레이아웃 거리가 커지진 않는다. 웹 페이지 내에 있는 동영상 수에 따라서 텍스트가 동영상과 상관관계가 있을 수도 있고, 그렇지 않을 수도 있기 때문에 웹 페이지가 포함하는 동영상의 수는 키워드 추출에 있어서 중요한 분류 기준이 된다. 이렇게 웹 페이지가 포함하는 동영상 수와 동영상으로부터 분포된 텍스트 형식에 따라 4가지 타입으로 분류가 가능하다. TYPE1은 웹 페이지가 단일 동영상을 포함하고, 동영상 주변의 텍스트가 비교적 간단한 타입이고, TYPE2는 단일 동영상만을 포함하면서 동영상 주변의 텍스트가 복잡한 구조를 가지는 형태이다. TYPE3는 웹 페이지에 여러 동영상이 포함되어 있고 텍스트들의 분포가 간단한 형태이며, TYPE4는 여러 동영상을 포함하면서 텍스트가 복잡하게 분포한 형태이다. 직접 수집한 1,087개의 웹 동영상 페이지(2,462개의 동영상을 포함)를 대상으로 분석한 결과 동영상의 분포가 TYPE별로 유사한 수치를 기록하는 것으로 보아서, 이와 같은 방법으로 타입을 분류하는 것은 의미가 있다.

웹 페이지로부터 키워드를 추출하기 위해서는 타입 별로 동영상 주변의 텍스트 블록을 찾고, 그 블록

들로부터 타입 특성에 맞는 키워드 추출 알고리즘을 적용한다. 추출된 단어는 중요도에 따라 가중치가 부여되는데, 이 가중치는 단어가 속한 텍스트 블록의 가중치와 각 단어의 특성에 따른 가중치의 곱으로 구한다. 텍스트 블록의 가중치는 레이아웃 거리에 반비례하고, 각 단어의 특성에 따른 가중치는 기존의 웹 이미지 페이지의 키워드 추출에서 사용한 가중치 대상인 빈도수나 HTML 태그 정보 외에 추가적으로 웹 동영상 페이지의 의미적 특성인 동영상 상세요약정보의 포함여부를 적용한다. 웹 페이지에서 키워드를 추출하는 알고리즘은 입력으로 동영상을 가진 웹 페이지와 추출할 키워드 개수가 주어진다. 먼저 웹 페이지에서 동영상과 주변 텍스트 블록을 추출한 후, 웹 페이지의 타입에 따라 가중치를 조정한다. TYPE1의 경우 추출된 텍스트에 모두 동일한 중요도를 적용하며, TYPE2는 레이아웃 거리에 따라 가중치를 조정한다. TYPE3의 경우 동영상이 여러 개 포함 되어있기 때문에 주변 텍스트가 어떤 동영상과 상관관계가 있는지를 결정해주어야 한다. TYPE4의 경우에는 동영상으로부터 떨어진 거리와 텍스트가 어떤 동영상과 관계가 있는지를 동시에 고려한다. 끝으로 단어들의 빈도수와 HTML 태그, 동영상 상세정보 등의 가중치를 반영하여 최종적으로 가중치를 결정하고, 상위에 위치한 키워드를 추출한다.

본 논문에서 제안한 방법에 의한 키워드 추출 결과를 평가하기 위해 대표적인 웹 이미지 검색 시스템에서 사용된 키워드 추출 방법 중 한가지인 ImageRover방식을 대상으로 재현율을 평가척도로 하여 비교하였다. 실험을 위해 사용한 데이터는 본 논문에서 구분한 웹 동영상 페이지의 4가지 타입 각각 30개로 이루어졌다. 실험결과 TYPE1은 기존의 키워드 추출 방법과 거의 동일한 구조이므로 제안한 방법에 대한 성능이 기존 방법과 거의 동일하거나 약간 낮게 나타났다. 웹 페이지의 구조가 TYPE2에 속할 경우에는 추출 키워드가 많아질수록 제안한 방법이 ImageRover에 비해 뛰어난 성능을 나타냄을 알 수 있다. 이는 ImageRover가 웹 페이지의 구조적인 특성은 고려하지 않고, 텍스트 상의 거리만을 평가하여 키워드를 추출했기 때문이다. TYPE3는 다수개의 동영상을 포함하는 간단한 구조의 웹 동영상 페이지로 구성되어 있어서 추출 과정에서 다른 동영상과 관련 있는 텍스트라고 판단되는 부분에서는 키워드를 추출 하지 않게 된다. 따라서 추출 되어야 하는 키워드가 의도하지 않게 필터링 되는 경우가 생길 수 있다. 따라서 이러한 이유로 TYPE3에 대한 키워드 추출 재현율이 추출 개수 7이하에 대하여 ImageRover보다 조금 낮게 나타난다고 할 수 있다. 마지막으로 TYPE4는 페이지 내 포함된 동영상의 수도 많고, 구성방식도 복잡하다. 따라서 ImageRover는 이러한 특성들을 고려하지 않기 때문에 재현율이 제안한 방법에 비하여 전체적으로 떨어지게 나타났다. 실험결과 본 논문에서 제안한 방법이 ImageRover와 비교하여 전체적으로는 재현율에 대해 18%의 성능 향상(키워드 8개 기준)을 보였다. 그리고 타입별 실험을 통하여, 제안한 방법이 간단한 웹 페이지 구성 방법인 TYPE1의 경우를 제외하고는 모든 타입에서 ImageRover와 비교하여 성능향상을 얻었으며, 특히 복잡한 형태의 웹 페이지 구조인 TYPE2, TYPE4에서는 ImageRover와 비교하였을 때 높은 성능평가를 보였다. 따라서 웹 동영상 페이지의 타입을 분류하고, 타입별로 키워드 추출방법을 달리하여 추출하는 것은 의미가 있다고 할 수 있다.

웹 이미지와 동영상 페이지를 분석한 결과 웹 페이지에서 이미지와 동영상이 차지하는 의미가 다르고, 페이지 구성방식에도 차이를 보이기 때문에 웹 동영상 검색에서는 기존의 웹 이미지에서 사용한 방법과는 다른 방법이 요구된다. 따라서 본 논문에서는 웹 동영상 페이지의 동영상과 텍스트의 구성 방식을 분석하고, 페이지 구성에 영향을 주는 요소를 바탕으로 페이지 구조를 4가지 타입으로 분류하고, 타입 별 키워드 추출 알고리즘을 제안하였다. 실험을 통하여 평가하여 본 결과, 본 논문에서 제안한 방법이 기존 웹 이미지 키워드 추출 방법인 ImageRover에 비해서 실험대상(120개의 웹 동영상 페이지)에 대하여 18%의 (키워드 8개 기준) 성능 향상을 보였다. 특히, 복잡한 형태의 웹 페이지 구조인 TYPE2, TYPE4에서는 ImageRover와 비교하였을 때 높은 성능 평가를 보였다. 따라서 웹 동영상 페이지의 타입을 분류하고, 타입 별로 키워드 추출방법을 달리하여 추출하는 본 논문의 제안 방법은 웹에 분포한 일반적인 웹 동영상 페이지로부터 키워드를 추출하는 방법으로 사용할 경우 기존의 방법에 비해 우수한 성능을 보일 것으로 예상된다.