자막 분석을 통한 교육 영상의 카테고리 분류 방안

이지훈 · 이현섭 · 김진덕*

동의대학교

Classification of Education Video by Subtitle Analysis

Ji-Hoon Lee · Hyeon Sup Lee · Jin-Deog Kim*

Dongeui University

E-mail : jdk@deu.ac.kr

요 약

본 논문은 교육 영상의 자막을 한글 형태소 분석기를 통해 추출하고 추출된 형태소 정보를 바탕으로 영상의 카테고리를 분류하는 방안에 대해 소개한다. 시스템에서 사람의 실수로 잘못된 정보가 입력되어 아이템의 특성으로 반영하게 될 경우 추천 시스템에서 정확도의 문제를 미치는 경우들이 있다. 이를 방지하기 위해 미리 분류된 영상에서 추출한 형태소 정보를 이용하여 각 카테고리에 해당하는 키워드 테이블을 생성하고, 각 카테고리 키워드 테이블과 영상의 형태소의 유사도를 비교하여 가장 유사도가 높은 키워드 테이블을 이용해 교육 영상의 카테고리를 분류한다. 이를 통해서 사람의 개입을 줄이고 시스템이 직접 영상을 분류하여 추천 시스템의 정확도를 높이는 것을 목표로 한다.

ABSTRACT

This paper introduces a method for extracting subtitles from lecture videos through a Korean morpheme analyzer and classifying video categories according to the extracted morpheme information. In some cases incorrect information is entered due to human error and reflected in the characteristics of the items, affecting the accuracy of the recommendation system. To prevent this, we generate a keyword table for each category using morpheme information extracted from pre-classified videos, and compare the similarity of morpheme in each category keyword table to classify categories of Lecture videos using the most similar keyword table. These human intervention reduction systems directly classify videos and aim to increase the accuracy of the system.

키워드

Lecture Video, Morpheme, Category, Subtitles, Corpus

Ⅰ. 서 론

기술의 발달로 인해 남녀노소 누구나 스마트 기기를 사용하고 있고, 인터넷의 기술이 발달하면서 언제 어디서든 다양한 영상을 쉽게 접할 수 있는 시대가 되었다. 그러면서 사용자가 이용할 수 있는 영상의 폭이 넓어지면서 너무 많은 영상으로 인해 원하는 영상을 선택하는 일에 어려움을 느끼기도 한다. 이 때문에 사용자 개인에게 맞춤형 콘텐츠를 추천하는 일이 중요한 요인으로 자리 잡았고, 대부분의 서비스의 핵심적인 부분 중 하나가 되었다.

추천 시스템의 중요도가 커지면서 아이템의 특성을 올바로 반영하여 사용자에게 더 정확한 추천을 할 수 있지만 사용자가 잘못된 정보를 입력할 경우 정확도에 큰 영향을 줄 수 있는 문제점이 있다. 이와 같은 경우를 방지하기 위해 사람의 개입을 줄여 시스템의 정확도를 높이는 것을 목표로 본 논문에서는 교육 영상의 특성 중 하나인 카테고리를 시스템이 스스로 분류하는 방안에 대해 기술하고자 한다.

본 논문에서는 정확한 카테고리 분류를 위해 영상의 내용, 즉 자막 정보를 이용하여 형태소를 분석하고 이를 바탕으로 영상에서 많이 사용되는 형태소를 추출하여 카테고리 키워드 테이블을 생성한다. 카테고리 키워드 테이블을 활용하여 신규 영상을 분류한다. 카테고리 키워드 테이블은 기존 영상 정보와 신규 영상 정보를 사용하여 새로운 정보를 반영하는 적응형 시스템이 될 수 있을 것으로 예상된다.

* corrresponding author

Ⅱ. 설 계

2.1. 시스템 설계

그림1은 카테고리가 분류되지 않는 영상을 분류하는 시스템의 구조를 보여주고 있다. 교육 영상을 분석하여 형태소를 추출하고 이를 관리한다. 영상의 랭킹정보를 바탕으로 영상을 선별하고 이 영상의 형태소 정보를 이용해 키워드 테이블을 생성한다. 생성된 키워드 테이블을 활용하여 신규 영상의 카테고리를 분류하도록 구성되어 있다.

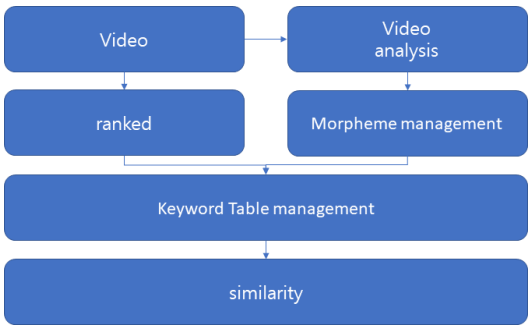

그림 1. 카테고리 분류 시스템의 구조

2.2. 영상분석

등록된 영상의 자막 정보를 이용하여 영상을 분석한다. 영상의 자막에서 형태소 분석기를 통해 형태소를 추출한다[1]. 카테고리 분류를 위한 키워드 테이블 생성을 위해서는 영상이 어떤 내용인지 판단할 필요성이 낮다. 또한 형태소를 다양하게 사용하기보다는 해당 카테고리에서 많이 사용되는 단어 정보를 이용하는 것으로 유의미한 결과를 얻을 수 있기 때문에 키워드 테이블을 생성할 때 명사에 해당하는 형태소만 남기고 나머지 형태소는 제거한다. 추출된 형태소에 불용어 사전을 이용하여 많이 사용되는 기본적인 형태소는 제거한다. 이런 과정을 통해 추출된 형태소 정보를 효율적으로 활용하기 위해 형태소의 빈도수를 계산하고 각 영상마다 분석한 정보를 빈도수 테이블 형태로 저장한다.

2.3. 키워드 테이블 생성 및 관리

키워드 테이블을 생성하기 위해 선별된 영상을 이용한다. 선별 방법은 해당 카테고리 영상 중에서 랭킹 수치가 높은, 조회 수나 추천 수 등의 수치가 높은 사용자에게 선호되는 영상을 이용한다. 기존 영상에서 70% 그리고 최근 1개월 동안 해당 카테고리에 새롭게 추가된 영상에서 30% 비율로 영상을 선별하여 이용한다. 최근에 추가된 신규 영상을 이용하여 새롭게 추가된 사실이나 변경 점을 반영하는 적응형 키워드 테이블을 생성한다.

선별된 기존 영상 그룹과 신규 영상 그룹의 형태소 테이블을 통합하여 키워드 테이블을 생성한다. 키워드 테이블 생성을 위해 사용될 영상의 수는 시스템의 환경에 따라 달라질 수 있지만, 영상의 수가 늘어날수록 키워드 테이블의 크기가 계속해서 커지기 때문에 테이블을 관리가 필요하다.

키워드 테이블을 관리하기 위해 형태소의 빈도수 정보를 랭킹화[2]하였다. 빈도수가 많은 순서대로 내림차순으로 정렬하여, 지정된 범위에 속하는 형태소만을 사용하여 키워드 테이블의 크기가 계속해서 커지지 않도록 방지한다.

빈도수가 높은 형태소를 이용할 경우 카테고리와 상관이 없이 한국어에서 많이 사용되는 형태소가 높은 빈도수를 가지는 경우가 있다. 이런 형태소의 경우 모든 영상에서 공통적으로 사용되기 때문에 유사도에 영향을 미친다. 이런 형태소들을 제거하기 위해 일반적으로 불용어 사전을 이용하여 불용어를 처리한다. 하지만 변화되는 영상의 특성을 반영하기 위해서는 사람의 개입이 주기적으로 필요할 수 있다. 이를 방지하기 위해 불용어 사전에 용어를 추가하지 않고 불용어를 제거한다.

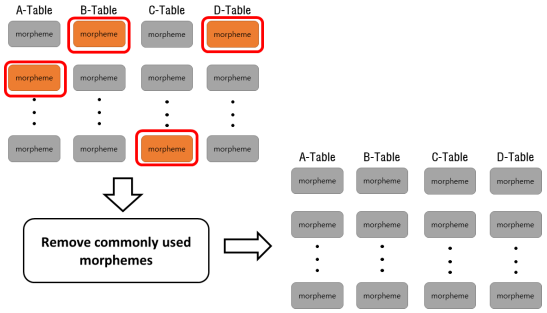

그림 2. 키워드 테이블 불용어 제거 방법

그림2는 키워드 테이블에서 불용어를 제거하는 방법이다. 모든 키워드 테이블에서 공통적으로 등장하는 형태소를 제거함으로 불용어를 처리하는 방법이다. TF-IDF[3] 방법의 경우 빈도수가 높은 형태소는 중요도를 낮춰 불용어를 판별한다. 하지만 키워드 테이블의 경우, 해당 카테고리에서 많이 사용되는 단어를 활용하여 영상을 분류하기 때문에 이 방법은 적합하지 않았다. 모든 키워드 테이블에서 공통적으로 사용되는 형태소를 제거함으로 형태소의 빈도수 정보를 반영하면서 불용어를 제거한 키워드 테이블을 생성한다.

2.4. 신규 영상 분류

앞의 방법을 통해 생성된 키워드 테이블을 활용하여 아직 분류되지 않은 신규 영상을 분류한다.

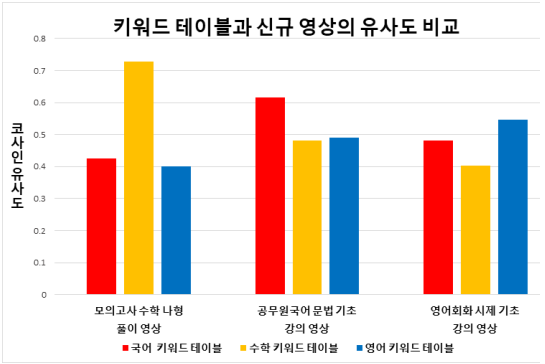

그림 3. 키워드 테이블과 신규 영상의 유사도 비교

신규 영상이 등록될 경우, 신규 영상을 분석하여 형태소 정보를 저장한다. 신규 영상의 벡터 테이블과 각 카테고리의 키워드 테이블을 비교하여 코사인 유사도[4]를 계산한다. 키워드 테이블과 신규 영상의 유사도 계산을 통해 유사도가 가장 높게 나타난 키워드 테이블의 정보를 이용하여 신규 영상의 카테고리를 지정한다.

그림3은 유튜브에 업로드된 교육 영상의 자막을 이용하여 만든 수학, 국어, 영어에 해당하는 키워드 테이블과 신규 영상의 유사도를 비교한 차트이다. 하나의 카테고리마다 약 50개의 영상을 사용하였으며 불용어 처리를 하지 않았다. 신규 영상의 경우 키워드 테이블의 생성에 사용되지 않은 교육 영상을 이용하였다.

각 차트의 붉은색은 국어, 주황은 수학, 파랑은 영어 키워드 테이블과 유사도를 의미하며, 각 신규 영상과 유사도를 비교했을 때 신규 수학 영상은 0.728로 수학 키워드 테이블과 가장 높은 유사도를 가지고 신규 국어 영상은 0.616으로 국어 키워드 테이블과 신규 영어 영상은 0.547로 영어 키워드 테이블과 가장 유사했다. 카테고리의 종류와 표본의 수가 부족하지만 유의미한 결과를 얻을 수 있었다.

Ⅲ. 결 론

본 논문은 교육 영상의 특성을 반영하는 일에 있어 사용자의 실수가 시스템에 정확성에 영향을 주는 경우를 방지하기 위해 사람의 개입을 줄여서 시스템의 정확성을 높이는 목적으로 수행되었다.

이 논문에서는 이미 카테고리가 분류되어 있는 영상의 자막 정보를 이용하여 신규 영상을 분류하는 것을 목표로 기존 영상의 자막을 형태소 분석기를 통해 형태소를 추출하고 이를 이용하여 각 카테고리에 해당하는 키워드 테이블을 생성한다. 생성된 키워드 테이블에서 모든 카테고리에 공통적으로 존재하는 형태소를 제거함으로 불용어를 처리하고, 빈도수 랭킹화를 통해 키워드 테이블을 효율적으로 관리하는 방안을 제안하였다.

이 과정을 통해 생성된 카테고리 키워드 테이블과 신규 영상의 유사도를 비교했을 때 해당 영상의 카테고리에 해당하는 키워드 테이블과 코사인 유사도 수치가 상대적으로 높은 유의미한 결과를 얻을 수 있었다. 이 점을 이용하여 신규 영상의 카테고리 분류가 가능할 것으로 예상된다.

시스템이 스스로 교육 영상을 분류함으로 사용자에게 편리함을 가져올 것이다. 이를 통해 좀 더 발전된 시스템을 사용자에게 제공할 수 있을 것이다.

References

[1] Hyoun-Sup Lee, Jun-Ho Kim, Jae-Chul Lee, Bo-Ah Na, Jin-Deog Kim, “Design of Word and Stemming Extraction System for Keyword Analysis”, Proceedings of the Korean Society for Information and Communication Sciences Conference 23(2), 2019.10, 538-539 (2 pages)

[2] Hyun-Sup Lee, Jindeog Kim, "A Design of Similar Video Recommendation System using Extracted Words in Big Data Cluster" Journal of Academic Presentation of the Korean Society of Information Sciences, Vol. 24, No.2, (2020): 172-178.

[3] Dae-Seo Park ,and Hwa-Jong Kim. "A Proposal of oin Vector for Semantic Factor Reflection in TF-IDF Based Keyword Extraction." Journal of the Korea Society of Information Technology, Vol. 16, o.2 (2018): 1-16.

[4] Minjae Kim, Sangjin Lee, "Measures of Abnormal User Activities in Online Comments Based on Cosine Similarity," Journal of KIISE, Vol.24, No.2 (2014):335-343.