The Journal of the Convergence on Culture Technology (JCCT)
Vol. 10, No. 2, pp.499-503, March 31, 2024. pISSN 2384-0358, eISSN 2384-0366

http://dx.doi.org/10.17703/JCCT.2024.10.2.499
JCCT 2024-3-58

AI를 활용한 메타데이터 추출 및 웹서비스용 메타데이터 고도화 연구

Metadata extraction using AI and advanced metadata research for web services

박성환*

Sung Hwan Park*

요 약 방송 프로그램은 자체 방송 송출 외에도 인터넷 다시 보기, OTT, IPTV 서비스 등 다양한 매체에 제공되고 있다. 이 경우 콘텐츠 특성을 잘 나타내는 검색용 키워드 제공은 필수적이다. 방송사에서는 제작 단계, 아카이브 단계 등에서 주요 키워드를 수동으로 입력하는 방법을 주로 사용한다. 이 방식은 양적으로는 핵심 메타데이터 확보에 부족하고, 내용 면에서도 타 매체 서비스에서 콘텐츠 추천과 검색에 한계를 드러낸다. 본 연구는 EBS에서 개발한 DTV 자막방송 서버를 통해 사전 아카이빙 된 폐쇄형 자막 데이터를 활용하여 다수의 메타데이터를 확보하는 방법을 구현했다. 먼저 구글의 자연어 처리 AI 기술을 적용하여 핵심 메타데이터를 자동으로 추출하였다. 다음 단계는 핵심 연구 내용으로 우선순위와 콘텐츠 특성을 반영하여 핵심 메타데이터를 찾는 방법을 제안한다. 차별화된 메타데이터 가중치를 구하는 기술로는 TF-IDF 계산법을 응용하여 중요도를 분류했다. 실험 결과 성공적인 가중치 데이터를 얻었다. 이 연구로 확보한 문자열 메타데이터는 추후 문자열 유사도 측정 연구와 결합하면 타 매체에 제공하는 콘텐츠 서비스에서 정교한 콘텐츠 추천용 메타데이터를 확보하는 기반이 된다.

주요어 : 구글 AI, 자연어 처리, 메타데이터 추출, 콘텐츠 서비스, TF-IDF

Abstract Broadcasting programs are provided to various media such as Internet replay, OTT, and IPTV services as well as self-broadcasting. In this case, it is very important to provide keywords for search that represent the characteristics of the content well. Broadcasters mainly use the method of manually entering key keywords in the production process and the archive process. This method is insufficient in terms of quantity to secure core metadata, and also reveals limitations in recommending and using content in other media services. This study supports securing a large number of metadata by utilizing closed caption data pre-archived through the DTV closed captioning server developed in EBS. First, core metadata was automatically extracted by applying Google's natural language AI technology. The next step is to propose a method of finding core metadata by reflecting priorities and content characteristics as core research contents. As a technology to obtain differentiated metadata weights, the importance was classified by applying the TF-IDF calculation method. Successful weight data were obtained as a result of the experiment. The string metadata obtained by this study, when combined with future string similarity measurement studies, becomes the basis for securing sophisticated content recommendation metadata from content services provided to other media.

Key words : Google AI, Natural Language AI, Metadata extraction, Content Service, TF-IDF

*정회원 한국교육방송공사 수석연구위원(제1저자) (교신저자)
(광운대학교 대학원 콘텐츠학박사)
접수일: 2024년 1월 2일, 수정완료일: 2024년 1월 23일
게재확정일: 2024년 1월 31일

Received: January 2, 2024 / Revised: January 23, 2024
Accepted: January 31, 2024
*Corresponding Author: parkslab@ebs.co.kr
Dept. of Sound Tech, EBS, Korea

Ⅰ. 서 론

방송사에서 제작한 방송용 프로그램은 자체 채널의 방송 서비스 이외에도 다양한 매체, 채널에 공급되어 소비되고 있다. 이때 콘텐츠 내용 검색을 위한 키워드 제공은 필수적이다. 이처럼 디지털 환경에서 콘텐츠의 유통, 검색 등에서 메타데이터의 중요성은 어느 때 보다 높다 [1]. VOD 서비스 뿐 만 아니라, 유튜브와 같은 OTT 플랫폼 서비스에서도 콘텐츠 메타데이터는 사용자 선호도를 반영한 추천도 중요하다. 하지만 방송사의 전통적인 메타데이터 입력 방법은 매우 비효율적이었다. 프로그램 제작 단계에 제작진이 입력하거나, 방송된 프로그램의 아카이빙 단계에서 담당자가 입력하는 방법이다. 이는 핵심 키워드 선정이 부정확하고, 시간과 인력을 투입해도 유용한 메타데이터를 다량 확보하기 어려운 한계점이 있다. 본 연구는 다양한 플랫폼에서 콘텐츠 서비스 이용률을 높이기 위해서 콘텐츠 특성에 따른 핵심 메타데이터의 선정, 콘텐츠 맥락과 소비자의 선호도를 반영한 양질의 메타데이터를 확보하는 것이다.

본 연구에서는 먼저 다수의 데이터를 확보하기 위하여 DTV 방송의 부가 서비스 중 청각 장애인을 위한 자막방송용 DTVCC(DTV Closed Caption) 텍스트 데이터에 구글의 상용화된 AI 기술을 적용하여 메타데이터를 추출하는 방법을 실험했다. 인공지능을 활용하면 영상 인식, 음성 인식, 문자 데이터 활용 방식 등으로 키워드 추출은 가능하다. 하지만 프로그램 특성을 반영한 중요 메타데이터 선별은 어려운 것이 현실이다.

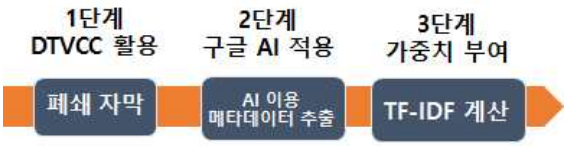

그림 1. 메타데이터 추출에서 가중치 계산까지 절차
Figure 1. Procedures from extracting metadata to recommending content

자연어 처리 인공지능을 활용하여 추출한 데이터가 단순 빈도수 중심이라는 한계를 가지므로 추가 연구로 메타데이터 가중치를 통한 우선순위, 중요도를 계산하는 실험을 실시하였다. TF-IDF 계산 방식을 적용하여 데이터 순위값으로 핵심 메타데이터를 확보하였다. 이렇게 확보한 메타데이터는 방송 프로그램명, 부제, 핵심 키워드와 결합하여 프로그램 상세 데이터를 제공하는 메타데이터 고도화 과정이라고 하겠다. 특히 이 데이터는 문자열 구조를 가지므로, 문자열 유사도 측정 연구와 결합하기 쉬운 구조이다. 코사인 유사도, 유클리디안 유사도, 맨하탄 유사도, 자카드 유사도, 오픈소스 검색엔진에서 사용하는 BM25 유사도 산출 알고리즘을 통한 방식 등과 접목이 용이하다는 장점이 있다 [2, 3]. 후속 연구로 유사도 측정 방식을 접목가면 콘텐츠 유통 서비스에서 콘텐츠 추천과 검색 품질 향상에 기여할 수 있다.

Ⅱ. 메타데이터 추출 자동화

1. DTV 방송 자막 데이터 활용

ATSC 방식의 HDTV 방송 시스템에서 청각 장애인을 위한 자막방송 서비스 표준은 EIA-708-D에 따른다 [4]. 자막 데이터 통신 방식 규격으로는 두 가지 방식이 있다. HD 영상, 음성 신호와 분리된 자막 데이터를 Video Encoder에 RS-232 신호로 입력하는 SMPTE-333M 규격 방식과 HD 신호에 자막 신호를 포함하여 HD-SDI로 전송하는 SMPTE-334M 규격이 그것이다 [5]. 이후 인코더에서 MPEG-2 TS(Transport Stream)로 다중화 되어 서비스 된다 [6]. EBS에서 자체 개발한 자막서버 시스템은 Video Encoder 특성을 고려하여 SMPTE-333M 규격을 준수하고 있다. 아래 (그림 2)와 같이 자체 개발한 자막 서버의 자막 데이터는 방송 후 아카이브에 저장되며 저장된 SubRip 자막파일에서 메타데이터를 추출하는 방식을 사용하였다 [7].

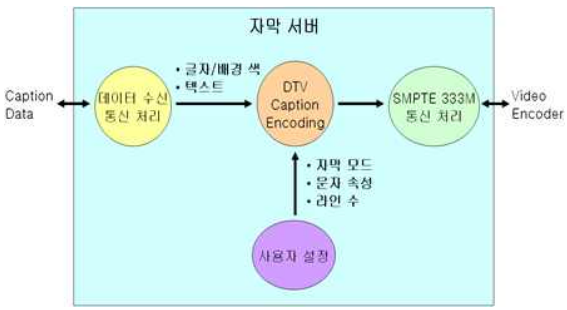

그림 2. DTV 자막방송용으로 개발한 자막서버 블록도
Figure 2. A caption server block developed for DTV closed captioning

2. 방송용 메타데이터 입력체계와 자막 데이터

방송국의 주요 키워드 입력 단계로는 프로그램 기획 단계에서 ERP 시스템에 주요 키워드 5개를 입력한다. 다음으로 부조정실, 편집과정의 비선형 제작단계에서는 내용 관련 메타데이터를 개수 제한 없이 입력 가능하지만 제작 스케쥴상 추가 입력은 어렵다. 그래서 기타 콘텐츠 서비스 플랫폼에 필요한 메타데이터를 웹서비스, 미디어 통합관리시스템에서 별도 입력하고 있다. 이 방식의 한계는 입력한 메타데이터가 E-DAS라는 디지털 아카이브 시스템에 저장되지 않는다. 결국 프로그램마다 프로그램 제목과 부제를 제외하고 보통 5개의 주요 키워드가 아카이빙 되어 있는 한계가 있다.

이외에도 외부 콘텐츠 유통을 위한 메타데이터 제공에도 15개 정도의 주요 키워드를 수동으로 입력하는 한계로 외부 플랫폼에서 검색 효율성이 떨어진다.

3. 메타데이터 자동 추출 결과

앞 절에 언급한 입력 메타데이터 부족을 극복하고 방송 프로그램에서 풍부한 메타 데이터를 추출하는 방식으로는 영상 인식, 음성 인식, 콘티와 같은 텍스트 활용 등이 가능하다. 본 연구는 정확도와 방송국 내부 기존 데이터 활용 측면에서 청각 장애인 방송에 사용된 자막방송 데이터 아카이브 자료를 활용 하였다.

그림 3. E-DAS에서 자막방송 파일 활용하기
Figure 3. Using Caption Files in E-DAS

E-DAS(EBS Digital Archive System) 시스템에 저장된 비정형 텍스트에 구글의 자연어처리 인공지능(Natural Language AI) API를 연계하여 유용한 키워드를 도출하였다. 구글의 자연어 처리 API는 감정 분석, 항목 분석, 항목 감정 분석, 콘텐츠 분류 및 구문 분석 등의 모델을 제공한다 [8]. 본 연구에서는 텍스트 블록에 적용되는 콘텐츠 카테고리를 식별하는 콘텐츠 분류 유형을 적용하여 프로그램 특성에 맞는 메타데이터를 추출하였다.

1차적으로 구글 자연처리 AI를 적용을 통해서 얻은 결과는 다량의 메타데이터를 추출할수록 정확도를 높일 수 있다는 것이다. 실제 실험결과 구글 자연처리 AI를 활용하면 프로그램 당 수백 개의 메타데이터를 자동 추출할 수 있다. 표 1의 예시에서는 사람이 입력한 데이터와 AI를 활용한 추출 수의 개수와 빈도수 차이를 알 수 있다.

표 1. 입력 데이터와 자막방송에서 추출한 메타데이터 비교
Table1.Comparison of input data with metadata extracted from closed captioning

	분류	1	2	3	4	5	6	7	8	9	10
사람	DAS	레드불	최원준	무신사							
사람	ERP	최원준	레드불	브랜드	콘텐츠						
자막	AI추출	콘텐츠	레드불	무신사	발견	얘기	콘텐츠 마케팅	이야기	돈	마케팅	접근

	분류	11	12	13	14	15	16	17	18	19	20
사람	DAS										
사람	ERP										
자막	AI추출	방식	광고	맛	브랜드	도전	자체	패션 콘텐츠들	물건	연결	유튜브

AI를 활용한 메타데이터 자동 추출은 작업 시간을 획기적으로 단축하여 경제성을 높였으며 데이터량은 20배 이상 확보가 가능하였다.

III. 메타데이터 가중치 적용 연구

AI를 활용한 텍스트 기반 메타데이터 자동 추출 방식으로 확보한 메타데이터는 출현 빈도에 따른 데이터로서 실제 프로그램 내용 기반의 추천용 중요도 측면에서는 한계를 가진다. 인터넷, OTT, 소셜미디어 등 다양한 플랫폼 서비스를 위해서는 효율적인 데이터 관리가 필요하다 [9].

1. 가중치 적용 메타데이터 생성방법

빈도 수 기반의 메타데이터에서 추천 메타데이터를 얻기 위한 프로그램 핵심어를 도출하는 방법으로는 텍스트 마이닝으로 사용하는 TF-IDF (Term Frequency -

Inverse Document Frequency) 기법을 적용하였다. TF-IDF는 단어별 가중치로 문서의 특징을 표현하여 두 문서 간 유사도를 비교하거나 문서의 핵심어를 추출하는 방법이다 [10, 11]. 세계테마기행 프로그램의 경우, 반복 등장하는 일상어인 '마을', '도시', '사람들' 등의 일반적인 단어들의 중요도를 낮추고, '파묵칼레', '휘파람', '쿠스코' 등 콘텐츠의 특징을 잘 드러내는 단어의 중요도를 높이는 계산법이다.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

본 연구에서는 EBR EBS 비즈니스 리뷰, 명의, 세계테마기행, 다큐프라임, 앙코르 다큐프라임 등 5개 프로그램에 총 6500여 편에 가중치 실험을 시행하였다. 방대한 데이터 처리를 위한 TF-IDF 알고리즘은 파이썬으로 구현하였다.

표 2. TF-IDF 계산 적용 사례
Table 2. TF-IDF Calculation Sample

키워드	TF	IDF	TF-IDF
음식	20	$\log(\frac{100}{90}) = 0.046$	0.92
이스탄불	2	$\log(\frac{100}{5}) = 1.3$	2.6

제시한 사례는 세계테마기행 프로그램 100편에 대한 자막 100편을 사용하였으며 그 중 "1만년 역사의 땅, 터키 -4부 터키와 코레" 라는 부제를 가진 프로그램에 자연어 처리 AI로 추출한 데이터는 자막 내에 "음식"이라는 단어가 20번 등장하고, "이스탄불"은 2번 등장하였다. 세계테마기행 100편으로 확대해보면 "음식"이 등장하는 자막은 90개, "이스탄불"이 등장하는 자막은 5개였다. 그래서 샘플 프로그램에서 TF-IDF 계산 결과는 "음식"이 0.92이고 "이스탄불"이 2.6으로 중요한 가중치를 가진다.

Ⅳ. 토 론

방송국의 전통적인 방송용 메타데이터는 제목, 부제, 방송일, 채널명, 콘텐츠ID, 프로그램 길이 등으로 방송 프로그램 제작과 송출을 중심으로 수동 입력하는 현실이었다. 그래서 씬이나 내용 관련 메타데이터는 프로그램 당 5개 정도로 다양한 매체에 방송 콘텐츠 서비스용으로 제공하기에는 매우 부족하다. 그래서 온라인, OTT 서비스 등에서 검색 및 콘텐츠 추천에 미비한 것이 현실이다. 콘텐츠를 노출시키기 위한 검색(Search), 추천(Recommendation), 탐색(Discovery)의 3가지 행동에 가장 필요한 필수 데이터가 부족한 것이다 [12].

이에 방송사의 메타데이터 입력 단계별 시스템을 살펴보고 2단계로 고도화 방안을 연구하였다. 1단계는 청각장애인을 위해서 제공하는 폐쇄자막 데이터에서 제공하는 방송 프로그램 출연자의 대화, 화면의 소리, 분위기를 자막으로 제공하는 데이터를 활용하였다. HD 방송용으로 개발한 자막방송 시스템의 자막 방송용 서버를 활용하여 아카이빙 된 확장자 Srt인 SubRip 자막 파일에서 다량의 메타데이터를 추출하였다. 적용 기술은 구글의 자연어처리 AI를 활용하여 프로그램 1편당 수백 개의 메타데이터를 자동 추출하는 것이 가능하다.

이렇게 확보한 데이터는 프로그램 내에 등장하는 빈도 중심의 데이터로 수량에 비해서 효용성은 부족하다. "최고의 요리비결"이라는 프로그램에 "요리"라는 단어가 가장 많이 나오고, "여행" 프로그램에 "여행"이라는 단어가 빈출 단어이기 때문이다.

프로그램 내용별 특성과 중요도를 고려하여 데이터의 우선순위를 정할 필요가 있다. 2단계로 가중치 메타데이터를 얻기 위한 방법으로 TF-IDF 계산법을 적용하였다.

EBS의 5대 인기 프로그램의 총 6500여 편에 가중치 실험을 진행하였으며, 세계테마기행 프로그램의 중요도 반영 메타데이터를 (그림 3)의 결과와 같이 E-DAS(EBS Digital Archive System)에 AI 메타 항목을 적용하여 활용도를 높이고 있다.

(그림 3)의 좌측에 보이는 LOCATION, EVENT, PERSON 등의 카테고리 추출은 구글 자연어 처리 AI에서 추출한 결과를 카테고리이다. 이러한 카테고리별 분류

항목 아래에 세부 메타데이터는 TF-IDF 계산 공식을 활용한 가중치를 반영하여 적용하였다.

그림 3 AI 메타데이터 추출 및 가중치 결과 적용
Figure 3. AI metadata extraction and weighting results applied

연구 결과물인 AI 메타데이터는 더 많은 실험을 통해서 더 정교하게 업데이트 될 수 있도록 개방적이다. 이후에는 방송 프로그램 제작 단계부터 생성, 편집이 가능한 메타데이터 워크플로우를 구축하는 방법으로 확장도 가능하다 [13].

V. 결 론

본 연구로 얻은 가중치 메타데이터는 콘텐츠 유통, 판매에 적용하여 효율을 높였다는 점, 타 매체, 여타 플랫폼에 콘텐츠를 제공할 때 매우 유용한 콘텐츠 추천 서비스의 추가 개발에 중요한 데이터를 확보했다는 점, 그리고 보유한 자막방송 데이터를 활용하여 시간적, 경제적으로 효율성을 높인 점에서 의미가 크다.

본 연구는 자막방송 텍스트 데이터를 기반으로 하였으므로 추후 문자열 유사도 측정 연구와 결합하기 쉽다. 콘텐츠 추천을 위한 메타데이터를 얻기 위해서는 코사인 유사도 이외에도 유클리디안 유사도, 맨하탄 유사도 등을 활용할 수 있다. 단기적으로는 메타데이터 추출 자동화와 중요도 선정을 기반으로 인터넷 서비스, OTT 서비스 등 외부 플랫폼에 제공하는 메타데이터의 완성도를 높이는 역할이 기대된다. 콘텐츠 추천과 검색의 품질 향상에 기여할 수 있는 후속 연구에 활용 예정이다.

References

[1] J.H. Kim, "User Experience Analysis of OTT Service Content Recommendation -Focused on Netflix Case", Journal of Integrated Design Research, Vol. 20, No. 2, pp. 77, 2021.
[2] TTA Standard, "Data Search system Using Metada-Based Ranking Algorithm", TTAK.KO-07.0093/R4, 2023.
[3] C.G. Hwang, "Sentence Similarity Analysis using Ontology Based on Cosine Similarity", KIICE, Vol 25, No. 1, pp. 441, 2021.
[4] W.Y. Choi, "Assistive Broadcasting Services for the Vision and Hearing Impaired", JBE, Vol. 27, No. 4, pp. 588, 2022.
[5] M.H. Kim, "Implement closed captioning systems for the deaf", Journal of Korea Game of Society, Vol. 16, No. 1. pp. 105, 2016.
[6] J.Y. Kim, "A Study on Multimedia Application Service using DTV Closed Caption Data", JBE, Vol. 14, No. 4, pp. 489, 2009.
[7] J.H. Song, "Development of the Closed-caption Broadcasting System", EBS Technology Research Institute, Vol. 8, pp. 95-133, 2007.
[8] Natural Language API, https://cloud.google.com/natural-language/?hl=ko
[9] S.H. Park, "A proposal for a UHD/S3D-intgrated media asset management architecture based on the analysis of the practical archiving system", The Graduate School of Kwangwon University, A Ph.D degree Thesis, pp. 3, 2016.
[10] E.S. You, G.H. Choi, S.H. Kim. "Study on Extraction of Keywords Using TF-IDF and Text Structure of Novels" Journal of the Korea Society of Computer and Information, Vol. 20, pp.121 - 129, 2015.
[11] QAISER Shahzad, ALI Ramsha. "Text mining: use of TF-IDF to examine the relevance of words to documents." International Journal of Computer Applications, Vol. 181, pp.25-29, 2018.
[12] J.H. Kim, "Realized AI and Synergy of broadcast content", KCA, Media Issue & Trend Vol. 52, pp. 66, 2022.
[13] Y.H Oh, "MXF-based Broadcast Metadata Authoring and Browsing", JBE, Vol. 1, No. 3, pp. 278, 2006.