Research on Educational Video Retrieval Method Based on Audio Transcription Technology

Muqiang Zhao[1,2, a], Wenxi Zheng[1, b], Yan Ye[1, c], Min Wu[1,2, d]

[1]Lab of Modern Educational Technology, University of Science and Technology of China, Hefei, Anhui 230026, China;

[2]School of Software Engineering, University of Science and Technology of China, Hefei, Anhui 230026, China

[a] zmq781279839@163.com, [b] wxzheng@ustc.edu.cn, [c] zyyp@ustc.edu.cn, [d] minwu@ustc.edu.cn

Keywords: Video retrieval, Audio transcription technology, Retrieve granularity.

Abstract. In the modern education platform, there are many problems, such as insufficient retrieval accuracy and single retrieval process, this paper conducts an in-depth investigation and analysis of the features of educational videos, combined with the cutting-edge audio transcription technology and text processing ideas, and following the policy of modularization, measurable and easy to expand the characteristics of the Internet industry. Finally, this paper successfully designs a retrieval method based on audio transcription technology of educational video. In order to achieve this goal, a detailed needs analysis is applied, successfully modularize the search process, reduce the retrieve granularity and improve the retrieval accuracy of educational video.

1. Introduction

In the modern internet environment, the education industry has developed rapidly, various advanced technologies have been used in the education industry. On the one hand, educational resource websites constructed using educational videos have appeared in the industry. On the other hand, many educational platforms such as MOOC also have a large number of educational video resources for learners to use. However, how to quickly and accurately retrieve the knowledge points of learners in many video resources such as educational websites and MOOCs has become an important topic for the efficient use of educational videos. The application of the latest audio transcription technology to video retrieval has received extensive attention [1]. This paper attempts to propose and explore a fast and efficient retrieval method based on audio transcription technology education video.

Video retrieval in the Internet industry can be divided into two types: word segmentation fuzzy query and sub-word meaning query. Through investigation, it has been found that the well-known websites that contain educational videos in China are fuzzy word-segment queries on titles and profiles, and individual foreign websites have applied word segmentation and semantic queries. In the era of multimedia information diversification, the limitations of the retrieval method have become increasingly evident. During the class, students can’t learn a large number of knowledge points effectively, and most students have the problem of weak self-study ability and inertia. Reducing the granularity of retrieval videos is one of the effective ways to solve these problems. In recent years, the typical video retrieval system based on audio classification is more and more [2]. The system is based on the classification of sound tracks in the environment and establishes the relationship between ambient sound and video. The processing process is complex and cannot be used in the Internet industry, but the research direction of using audio to retrieve video has been widely adopted.

The typical processing of audio transcription technology is feature extraction, audio segmentation and audio word conversion. The feature extraction is based on some theoretical studies of sound correlation to select feature values. The selection of feature values mainly refers to the time domain, frequency domain, time-frequency domain and cepstrum analysis mentioned in audio information technology [3]. At present, the audio transcription technology based on deep full-sequence convolutional neural network can be used for industrialization in terms of processing cost and precision. We looked through research that the use of this theory to achieve the excellent performance of audio transfer products. In the video retrieval method in this article, the audio transfer process chooses IFlyTek voice transfer technology. The accuracy rate, recall rate and retrieval speed are commonly used in video retrieval [4]. The evaluation of retrieval results in this paper also depends on these indicators.

2. Video Retrieval Method

The video retrieval method process is divided into several stages: video preprocessing, audio transcription, text reorganization, word segmentation, filtering, and inverted storage. Video preprocessing is a process of processing video files into audio corresponding to a specific format. In this paper, the unified audio format is a sampling rate of 16k, and the channel is a mono wav. The pre-processing process uses FFmpeg to process the video to get the target audio; the audio transcription is to use IFlyTek audio transfer interface. This article has measured the effect of different types of audio transfer. The detailed experimental data shown in next section, where the results show that, for some relatively standard Chinese curriculum videos, the transfer interface is sufficient to support the formation of search data; the purpose of text reorganization is to divide the audio corresponding to a video into several segments. The length of each segment is selected for ten minutes, and this article gives the reason for choosing ten minutes; the word segmentation process is to split the Chinese-British mixed text corresponding to each audio segment, then process results and perform word frequency statistics; filtering has two types: filtering non-professional vocabulary and text and filtering professional vocabulary. The reason is to reduce the retrieval time and improve the retrieval efficiency; Inverted storage is to filter the search words into the database, the use of inverted index storage mode, characterized by a simple search and easy to organize the search results. The professional vocabulary used for filtering professional vocabulary comes from the teaching materials. Using this word set, the video can be automatically tagged to improve the retrieval efficiency. Name these professional words for the tag pool. The audio video transcription technology based educational video retrieval method at each stage of the relationship is shown in Figure 1.

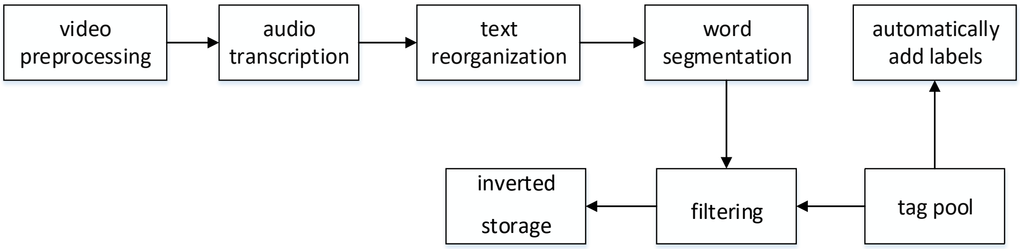

Fig. 1 Video retrieval methods at various stages of relationship

3. The Key Point of the Retrieval Method

3.1 Audio Transcription Technology Accuracy

In order to improve the measurement in the usability of the retrieval method, this section mainly contains the performance of audio transcription technology in the experiment. The experiments of audio transfer content in this paper are based on the audio transcription technology of IFlyTek voice recognition. For the classification of collections containing n audio corresponding documents, the number of each class is q, the total number of categories is k, and the audio corresponding document collection is:

$$\{\{D_{1,1}, D_{1,2}, \dots, D_{1,q}\}, \{D_{2,1}, D_{2,2}, \dots, D_{2,q}\}, \dots, \{D_{k,1}, D_{k,2}, \dots, D_{k,q}\}\} \tag{1}$$

One of the audio corresponding documents is named $D_{i,j}$, the number of words is named $DC_{i,j}$, the correct number of words after transfer is named $CC_{i,j}$, the average accuracy AC calculation formula:

$$\mathrm{AC} = \frac{1}{kq}\sum_{1\le \mathrm{i} \le \mathrm{k}}\sum_{1\le j\le q}\frac{CC_{i,j}}{DC_{i,j}} \quad (2)$$

Accuracy AC_i formula for each category:

$$\mathrm{AC}_i = \frac{1}{q}\sum_{1\le \mathrm{j} \le \mathrm{q}}\frac{CC_{i,j}}{DC_{i,j}} \quad (3)$$

The deviation of the category is the DF_i, formula:

$$\mathrm{DF}_i = |AC - AC_i| \quad (4)$$

Table 1. Accuracy and Deviation of Audio Transfer Results

Category	Accuracy	Offset	Average accuracy
Science and engineering	0.87	0.03	0.90
Liberal arts	0.92	0.02	

Judging from the experimental results, the performance of science and engineering is relatively weak, and the following some reasons are found through manual observation: when some of the proper nouns in science and engineering videos are mixed with Chinese and English, the transfer results are poor, the reason is that some teachers use dialect. The main reasons for the better liberal arts classes are that the nouns are mostly all Chinese, and the teacher's mandarin pronunciation is good, but there are also inaccuracies in the segmentation. The average accuracy is 90%. The official announcement rate of the official information of the company is up to 95%, but there are many influential factors in audio transcription, so it is lower than the speech recognition accuracy rate.

Overall, 90% of the accuracy rate has reached a searchable level. This paper develops process analysis and model construction based on this accuracy rate.

3.2 Chinese Word Segmentation

Chinese word segmentation is an important stage in the formation of retrivel terms. Chinese word segmentation contains two methods: mechanical matching methods and statistical methods. In this case, we use the probabilistic language model word segmentation method in statistical methods [5]. From a statistical point of view, the input to the word segmentation problem is a string $\mathrm{C} = C_1, C_2, \dots, C_n$, and the output is a word string $\mathrm{S} = W_1, W_2, \dots, W_m$, $\mathrm{m} \le \mathrm{n}$. A string C corresponds to multiple S. One of the output strings is S_i. The method aims to find the most likely segmentation result. Calculate each $\mathrm{P}(S_i|C)$ to find the largest one. For convenience calculation, according to the Bayesian formula:

$$\mathrm{P}(S|C) = \frac{P(C|S)\times P(S)}{P(C)} \quad (5)$$

Calculate the following formula:

$$\mathrm{Seg}(c) argmax_{S\in G} P(S|C) = argmax_{S\in G}\frac{P(C|S)P(S)}{P(C)} \quad (6)$$

$P(C)$ is the frequency of the string in the corpora, and it is a fixed value, $P(C|S)$ is the conversion probability from the string to the string, and the string is just one, so the probability is 1, The $\mathrm{P}(S_i|C)$ maximum can be converted to the maximum $P(S_i)$. We assume that the context is not related to words, and $P(S_i)$ is transformed into the product of the frequencies of the words in the string of words. Because the frequency of each word is independent, this calculation process satisfies the optimal sub-structure property and no-post-effect of dynamic programming. In order to reduce the amount of calculation, dynamic programming algorithm is used to calculate the segmentation scheme with the highest probability.

The segmentation corpus used in the Chinese word segmentation process does not include the special professional vocabulary related to the course. Therefore, in the word segmentation process, the original corpus is expanded using the professional vocabulary in the tag pool. For example, there is a term in the compiler theory course: “lexical analyzer”. The result of using no expanded corpus segmentation is:

{… , "lexical ", "analyzer", … }

One word in the tag pool is the lexical analyzer. The expanded corpus can be correctly obtained in experiments:

{… , "lexical analyzer", … }

3.3 Education Area Tag Pool

The field of education contains many professions, and each profession has many professional vocabularies, and these professional vocabularies form a pool of labels. In order to make the video retrieval method suitable for the education field, applying the tag pool to the Chinese word segmentation stage can effectively obtain the word segmentation result containing the professional vocabulary.

After investigation, it was found that there are many professional vocabularies on the final pages of most course materials. These courses save the manual tagging time, and do not require collecting professional vocabularies from textbooks, so, the professional vocabularies can be directly used in tag pools. The source of textbooks that do not have professional vocabulary at the end is the multimedia resources used by textbooks and teachers.

3.4 Retrieve Clip Time Selection

The narrowness of the video retrieval granularity in the retrieval method is reflected in the segmentation of the audio-transmitted document, and the selection of the segmentation segment time is based on the statistics of knowledge points. There are several main characteristics in the statistical process: In the science and engineering department's statistical process, one knowledge point only taught once in a 50-minute course, some knowledge points are short-term, and knowledge points are also not averaged; the use of liberal arts knowledge is relatively rare, and there are many examples of actual cases in the video, which are difficult to measure with time. We adopt a strategy to count according to the number of bars in the knowledge point and obtain the result:

Table 2. Relationship Between Video Duration and Bars

Category	Total video (hours)	Number of sections	Average time (minutes)
Science and engineering	21.3	115	11.11
Liberal arts	9.8	79	7.44

According to the statistical results of the number of sections within the knowledge point, the retrieval period is ten minutes. At this time, each section is not strictly divided, but the retrieval segment is divided according to the content of the video knowledge.

4. Retrieval Result Evaluation and Cost Analysis

Retrieval results evaluation indicators: precision, recall, and retrieval speed. Define N_c, N_f, N_m as the number of correct, incorrect, and missing words. The precision rate and recall rate are defined as follows:

$$\text{precision} = \frac{N_c}{N_c+N_f} \tag{7}$$

$$\text{recall} = \frac{N_c}{N_c+N_m} \tag{8}$$

The experimental results show that the precision and recall rate of the retrieval method in this paper are mainly affected by two factors: audio transcription accuracy and accuracy of Chinese word segmentation. The precision rate is 100%. The recall contains two experiments: before and after expansion of the corpus. Before the expansion, the average recall rate for arts and science education videos was 83%, and the expansion of corpus reached 91%. This experiment also verified the effectiveness of the expanded corpora. Because the retrieval method generates retrieval term set process and the user's retrieval behavior is asynchronous, so the retrieval speed of the retrieval method in this article depends on the database retrieval speed. After using the tag pool to filter the retrieval word set, the retrieval word set volume is small and the retrieval speed is low, which makes it qualified for actual use requirements.

The cost of the retrieval method is mainly in the stage of invoking the audio transcription technology. For the value of the high reuse rate education video, a certain cost is worthwhile. After the maturity of audio transcription technology in the future, the processing costs will decrease. However, the value of educational videos will change slowly, so the future trend of processing education videos will gradually appreciable.

5. Conclusion

This paper makes an in-depth research on educational video resources. According to the student's learning schedule and the demand for educational videos, an educational video retrieval method based on audio transcription is designed. It makes full use of advanced technology and applies it to practical video retrieval. Each stage of the retrieval method measures its effectiveness through experimental data, and a video retrieval function module based on this retrieval method is implemented and operated. The result shows the feasibility and innovation of the retrieval method.

References

[1]. Xueyuan Zhang, Qianhua He, Yanxiong Li, Yuling Ye. An audio retrieval method based on inverted index[J]. Journal of Electronics and Information Technology, 2012,34(11):2561-2567.

[2]. Issam Feki, Anis Ben Ammar, Adel M. Alimi.Automatic environmental sound concepts concepts for video retrieval[J]. International Journal of Multimedia Information Retrieval, 2016:114.

[3]. Jiqing Han, Tao Feng, Guibin Zheng. Audio Information Processing Technology [M]. Tsinghua University Press, 2007:32-46.

[4]. Yuxin Peng, Ngo Chong-Wah, Zongming Guo, Jianguo Xiao. Content-based video retrieval key technologies [J]. Computer Engineering, 2004:16.

[5]. Gang Luo, et al. Decryption Search Engine Technology Actual Lucene & Java Essence [M]. Electronic Industry Press, 2016: 159-172.