An agricultural audio retrieval method based on inverted index of silence word

Yue Song1,2,*, Sha Tao1,2,a, Yanzhao Ren1,2,b, Xinliang Liu1,2,c, Wanlin Gao1,2,d

1 College of Information and Electrical Engineering, China Agricultural University, 100083, Beijing, China
2 Key Laboratory of Agricultural Informationization Standardization, Ministry of Agriculture, Beijing 100083, China
*aCorresponding author e-mail: songyue597@163.com, taosha20070608@163.com, bxiaozhaochina@163.com, c569489372@qq.com, dgaowlin@cau.edu.cn

Abstract. With the popularization of Internet and multimedia technology, a large amount of agriculture-related audio information is rapidly accumulated and widely used. The method of keyword matching retrieval for text data content describing audio files can no longer meet the need of agricultural audio files retrieval in terms of retrieval efficiency and content depth. In view of the above problems in agricultural audio retrieval technology, this paper proposes an audio retrieval method based on inverted index of silence word. In this method, a new hybrid retrieval model combining representation retrieval and semantic retrieval is designed and used. The experimental results show that the recall rate and precision rate of this method are above 99.5%, and the average retrieval time is less than 1s. Therefore, this method greatly improves the efficiency and the accuracy of the audio information retrieval, and can be used for rapid retrieval of agricultural audio.

1. Introduction

In recent years, with the rapid development of Internet technology, more and more multimedia materials appear in daily life. And it has a great influence on people's study, work and family. Obviously, more and more agricultural technology audio is adopted by many agricultural workers, from which they can learn more and more agricultural knowledge. However, sometimes they can only enjoy part of the agricultural audio clips and it is difficult to find the original audio. Therefore, many scholars apply the method of text retrieval to audio retrieval. They are concerned about how to search the most ideal audio from the massive multimedia database quickly and accurately [1].

Audio retrieval can be divided into two parts that presentation retrieval and semantic retrieval [2]. The former takes the characteristics of sample waveform as the retrieval object and uses audio directly to retrieve audio without definition and training model in advance. However, this method queries all audios in the database, it consumes a lot of time in the face of massive database [3]. Semantic retrieval uses a priori model to retrieve audio data, which is suitable for mostly sound effect and audio scene [4, 5]. Its core idea of audio retrieval is based on speech recognition and text retrieval. However, speech recognition for a specific scene must be based on a professional lexicon. For instance, there are many professional words in agricultural popular science video. The effect of semantic retrieval applied to agricultural audio retrieval is not very good. A fast browsing search algorithm is proposed in reference [6], which is an improved representation retrieval method. Even though the retrieval speed of this method is greatly improved, the retrieval time is still long. Inverted index is widely used in text
retrieval, and mainly focuses on speech document retrieval in the field of audio retrieval [7]. It is applied to music retrieval [8] and birdsong sound retrieval [9]. Zhang uses inverted index method in text search engine to build audio dictionary and inverted index for audio, and proposes an audio retrieval method based on inverted index [3]. But the accuracy of agricultural audio retrieval is low in application.

Agricultural popular science video covers many fields, such as seed industry, planting industry, animal husbandry, aquaculture, etc. There are many special terms in each field, and different regional languages vary greatly. The audio of agricultural science video is mainly voice, and most of them are standard pure voice reading or voice with more soothing background music. In this paper, an agricultural audio retrieval method based on inverted index of silence word is proposed. First, multi-dimensional feature vector is built by many audio features. Second, a sequence of sequential audio segments is converted into a sequence of characterizing silent features. Then the inverted index is built based on silent feature sequence. Finally, this paper designs a hybrid retrieval model which combines representation retrieval and semantic retrieval to retrieve. At the beginning, the representation retrieval is used to save retrieval time, and then the semantic retrieval is used to improve retrieval accuracy.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed method. Section 4 discusses the experimental results and Section 5 concludes the paper.

2. Related work

In text retrieval, sequential retrieval means traversing every word of each article and judging whether it matches with query keywords. When the number of articles is large, the efficiency of this method is extremely low. Inverted index is the most commonly used index method in modern search engines [10]. It converts document-word information flow to item-document information [11], which mainly consists of term dictionary, term index and posting list.

Term dictionary is a string collection of all the terms that have appeared in a document collection. Each index entry in the term dictionary records some information of the term itself and a pointer to the inverted list.

Term index is a rule to find terms faster.

The posting list records the all documents list of a term and the location information of the document. Each record is called a posting. According to the posting list, user can know a term belong to which documents.

The diagram of inverted index can be showed in figure 1.

![Inverted Index](image)

Figure 1. Inverted Index.

Posting list is an array that stores all document ID that match the term. In the inverted index, user can find the position of the term in the term dictionary through the term index, and then find the posting list. With the posting list, user can find the document according to the ID.

3. The proposed method

Search engine must have two functions: index processing and query processing. The purpose of index processing is to establish a data structure that can be searched quickly. Query processing generates ordered search results based on these data structures and queries submitted by users.

3.1. Feature extraction

There are many kinds of audio features, which can be divided into different categories according to different rules. Each audio feature describes one aspect of the audio signal's characteristics. In this
paper, the 21-dimensional vector is used as a feature vector to fully represent the characteristics of various audio signals. It includes short time energy (STE), zero cross rate (ZCR) in time domain, spectrum energy (SE), spectrum centroid (SC), sub-band energy distribution (SED) Bandwidth (BW) in spectrum domain and 12-dimensional Mel Frequency Cepstrum Coefficients (MFCC) [12]. Because the numerical range of the features is very different, it is necessary to normalize the features. The normalized expression formula is shown in (1)

\[ x_i' = \frac{x_i - \mu_i}{\sigma_i} \]  

(1)

Where \( x_i \) is the \( i \)th original feature, \( \mu_i \) is the mean, and \( \sigma_i \) is the standard deviation. In this method, the features of each dimension are normalized to the distribution with the mean value of 0 and the variance of 1.

3.2. Silence Feature Sequence Construction

In multimedia data, silence is the most obvious segmentation point. The length of silence, distribution rule and order can be used as matching target. In this paper, the rule-based method is used to detect the silence segment, which is filtered by the dynamic threshold of STE. The frame whose short-term energy is lower than the threshold \( E_{th} \) is defined as the silence frame. \( E_{th} \) can be expressed as follows:

\[ E_{th} = E_{range} + E_{min} \]  

(2)

Where \( E_{range} \) represents the floating range of energy, and its value can be adjusted by the silence factor \( \lambda \).

\[ E_{range} = \lambda \cdot \min \left\{ \frac{1}{2} (E_{range} + E_{min}), E_{mean} - E_{min} \right\} \]  

(3)

Where \( E_{max} \) represents the maximum of audio frames, \( E_{min} \) represents the minimum of audio frames, \( E_{mean} \) represents the mean of audio frames. In this paper, \( \lambda \) is set to 0.002.

The isolated silence frame has no great significance and it may even be misjudged which can not be recognized as an effective feature. The silence feature is considered to appear \( L_{th} \) consecutive silence frames. According to the characteristics of agricultural popular science video, there are two kinds of silence frames: Silence Fragment (SF) and Silence Pause (SP). The former is the pause of the theme paragraph, and the latter is the pause of the speaker's reading. The frequency of SP is much higher than that of SF. In this paper, the length and timing information of SP are reconstructed to feature sequence. SP sequence (SS) extraction process is shown in Figure 2.

![Figure 2. SS extraction process.](image)

3.3. Build inverted index based on Silence Word (IBSW)

At present, some scholars apply inverted index for audio retrieval [13, 14]. They quantify the high-dimensional non-character data of audio to the form of some characters. In this paper, SS is quantized into silence word (SW) to realize the audio file character. To prevent large fluctuation of SS, directly discard SS at both ends of the sample. Only the middle complete SS is reserved as the later retrieval basis. The variance of SP is not more than 5 frames after processing, SW can be showed by the formula (4)
\[ SW = (SS - first - last) \mod 5 \] (4)

Where \( first \) and \( last \) represent the first and last SP in SS respectively.

In the traditional inverted index, the order between words is ignored. If only use a single word as the index, the method may ignore the timing sequence. In this paper, in order to make up for the lack of reflecting the timing characteristics of audio, IBSW uses two adjacent pairs of SW as the word item in the dictionary which needs to keep the separator. Index of IBSW dictionary should be:

\[ S'_1 = \{ (A, B), (B, C), (C, D), (D, E) \} \] (5)

Where \( A, B, C, D, E \) is silence word. Therefore, the 2-D array dictionary is used as the data structure of the dictionary. Each SW pair is regarded as a coordinate, and its position in the dictionary can be directly obtained by the term itself. The matrix mapping relationship is as follows:

\[
\text{Dictionary } [A-6][B-6]
\] (6)

Where \( A \) and \( B \) are two adjacent SW. The matrix size of the dictionary is \( D_{95 \times 95} \), which can meet the range of SW. Compared with the commonly method that is hash table, the structure of the dictionary using two-dimensional array does not need to match and search the terms. The retrieval efficiency of proposed method is improved. In the dictionary, each array element represents a word item which has two attributes: inverted document frequency (IDF) and inverted table pointer. The former and the term frequency (TF) information together determine the weighting factor. The latter points to the posting list corresponding to the term.

In this paper, SS is used as a word term which makes full use of the representation characteristics of speech. By using the random storage property of two-dimensional array, the corresponding posting list address can be found directly without term search and matching.

3.4. Hybrid retrieval model

The audio retrieval method which is based on IBSW only improves the retrieval efficiency. However, it does not improve the retrieval accuracy. In this paper, a hybrid retrieval model which is combined of IBSW and feature matching is proposed. In the initial stage of retrieval, IBSW is used to improve the efficiency of audio retrieval. Continue to retrieve the results from the above search, which uses feature matching which is based on 21 dimensional feature vector to improve retrieval accuracy.

In the retrieval process, the retrieved audio is first extracted by SW. When all candidate paragraphs are detected, they need to be sorted by TF-IDF. In this paper, TF-IDF mean value is used to reflect the possibility of target files. It is expressed by follows:

\[
p = \sum_{i=0}^{n} tf_i * idf_i
\] (7)

Where \( n \) is the number of detected inverted records relate to the audio file. The \( tf_i \), represents the term frequency of the \( i \)th inverted record. The \( idf_i \) represents the inverse document frequency of the \( i \)th posting list. Count the P value of all inverted records corresponding to each document. Only the audio files with higher P value continue to match subsequently. Therefore, this method realizes index removal and reduces the number of segments of feature matching which improves retrieval efficiency.

In this paper, segment feature which is 21-dimensional SW fragment feature (SWF) is extracted for the next 1s audio segment of each silence word. However, the segments at the end of the file may be not long (length < 1) enough to extract SWF. There may be three cases which are shown in figure 3. When the length of segments at the end of the file is less than 1s, we will discard the last SS directly.
The calculation of SWF similarity adopts Euclidean distance as the measure standard of similarity. The candidate audio file SWF is represented as $\mathbf{S}\mathbf{S}\mathbf{S}_i (x_1, x_2, x_3, \ldots, x_{21})$, query audio file SWF is represented as $\mathbf{S}\mathbf{S}\mathbf{S}_i' (y_1, y_2, y_3, \ldots, y_{21})$. The calculation formula is as follows:

$$\text{Sim}_i = -\sqrt{\sum_{k=1}^{21} (x_k - y_k)^2}$$

(8)

Calculate the similarity between query and all candidate documents. And according to the order from large to small, the ordered results are returned to the user to complete the retrieval process.

4. Experiments and Results

4.1. Dataset and Experiment Setup

The experimental data used in this paper are mainly from agricultural popular science audio, such as news broadcast programs and hundreds of forum programs etc. The audio content which covers many agricultural sub fields such as breeding, planting, machinery is extracted by Cool Edit Pro software. The total time of audio is about 100 hours as the retrieval source of the system. Before processing the data, it is necessary to use cool edit pro to extract the original data in mono channel. After reading by MATLAB, FS will be sampled down to 8kHz and pre weighted. Then add Hamming window for framing, and the frame length is 25ms. Through the “translate” function to pre-processing the audio and extract 21-dimensional features. Finally translate the original audio signal into SW sequence.

The specific experimental setup recall rate and precision rate is as follows: the test is performed when the audio files in the test set is 100, 200, 300, 400 and 500. The recall rate and the precision of the proposed method in this paper is compared the traditional sequential retrieval (TSR) method, Fibonacci hashing retrieval (FHR) method and middle fingerprint filtering retrieval (MFR) method.

The specific experimental setup retrieval time is as follows: randomly select 3000 audio files in the test database. And 500 agricultural audio clips of 20 s, 40 s, 60 s, 80 s, 100 s and 120 s are intercepted from these audio files. The retrieval time of the audio retrieval method proposed in this paper is compared with other three methods.

4.2. Evaluation Measures

Generally, there are three criteria to evaluate the retrieval methods in the conventional audio retrieval.

Retrieval accuracy (RA) refers to the number of similar audio retrieved and the ratio of all audio to video retrieved. The formula is as follows:

$$\text{RA} = A / (A + B)$$

(9)

Where A represents the number of similar audio retrieved, B represents the number of dissimilar audio retrieved.
Recall rate (RR) is used to describe the ratio between the number of recalled music pieces and the number of music pieces to be recalled. This index is widely used to evaluate the accuracy of information retrieval methods. The formula is as follows:

\[ \text{RT} = \frac{A}{A+C} \]  

(10)

Where C represents the number of audio not retrieved.

Another criterion is retrieval time (RT).

4.3. Experimental setup of recall rate and precision rate

According to the length of the bar graph, it can be seen that the recall rate of the proposed method in this paper is the highest among the four methods, and the mean value is above 99%. The figure 4 indicates that the information retrieval rate of this method is the highest.

![Figure 4. The recall rate of audio retrieval by four methods.](image-url)

In order to highlight the advantages of this method, 20% of comprehensive other information (not agriculture) is added to the above test to interfere with the retrieval. The above experiment is re-executed and the results are recorded.

![Figure 5. The recall rate of audio retrieval by four methods under 10% interference.](image-url)
Analysis of Figure 5 shows the comparison of the recall rates of the four methods with 10% interference information added. Overall, the proposed method is the highest among the four methods, indicating that the proposed method has the highest recall rate. It can be seen from the data that the precision of this method appears to be slightly fluctuating. Figure 5 shows that the audio retrieval of the proposed method is comprehensive and has advantages over other three methods.

Analysis of Figure 6 can clearly see the comparison of the values of the four methods. It is not difficult to see from the data in the figure that the precision rate of proposed method is higher than the other three methods. In summary, the proposed method in this paper can be used to digitize the effective retrieval of audio.

4.4. Experimental setup of efficiency

Analysis Table 3 shows that the time of the proposed method in this test is faster than other three method. Based on the data results in Table 3, it can be seen that the method of this paper retrieves audio in a shorter time and has the advantage of high efficiency.

| Length(s) | This paper prosed | TSR        | FHR        | MFR        |
|-----------|-------------------|------------|------------|------------|
| 20s       | 0.12              | 1008       | 123.57     | 56.78      |
| 40s       | 0.58              | 1523.5     | 200.35     | 32.36      |
| 60s       | 0.36              | 1353.21    | 400.54     | 45.21      |
| 80s       | 0.87              | 1645.35    | 325.89     | 30.11      |
| 100s      | 0.05              | 613.78     | 563.78     | 10.36      |
| 120s      | 0.15              | 1249.5     | 457.21     | 20.14      |

5. Conclusions

Limitations of retrieval methods of agricultural audio, the method for retrieving agricultural video based on inverted index of silence word is proposed. And a new hybrid retrieval model which combines representation retrieval and semantic retrieval is designed. At the beginning, the representation retrieval is used to save retrieval time, and then the semantic retrieval is used to improve retrieval accuracy. The experiment proves that the proposed method can obtain higher rates of recall and precision while spending less time and has the advantage of high efficiency. The method proposed in this paper not only makes up for the shortcomings of traditional methods, but also provides the effective and scientific retrieval method for audio, which provides the reference for agricultural audio processing.

References

[1] Heryanto H, Akbar S, and Sitohang B. Direct access in content-based audio information retrieval: a state of the art and challenges[C]. 2011 International Conference on Electrical Engineering and Informatics, Bandung, Indonesia, July 17-19, 2011: 1-6.
[2] Han, J. Q, Zheng Y. R., & and Zheng G. B. (2011). Theory and Technology of Audio Information Retrieval. Beijing: Science Press.
[3] Zhang X Y, He Q H, Li Y X, et al. An Inverted Index Based Audio Retrieval Method [J]. Dianzi Yu Xinxie Xuebao/Journal of Electronics and Information Technology, 2013, 34(11):2561-2567.
[4] Ghoraani B and Krishnan S. Time-frequency matrix feature extraction and classification of environmental audio signals[J].IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2197-2209.
[5] Fu Zhou-yu, Lu Guo-jun, Ting Kai-ming, et al. Music classification via the bag-of-features approach [J]. Pattern Recognition Letters, 2011, 32(14): 1768-1777.
[6] Kashino K, Kurozumi T, and Murase H. A quick search method for audio and video signals based on histogram pruning [J]. IEEE Transactions on Multimedia, 2003, 5(3): 348-357
[7] Matthews B, Chaudhari U, and Ramabhadran B. Fast audio search using vector space modeling[C].

7
IEEE Workshop on Automatic Speech Recognition & Understanding, Kyoto, Japan, Dec. 9-13, 2007: 641-646.

[8] Cha Guang-ho. An effective and efficient indexing scheme for audio fingerprinting[C]. 5th FTRA International Conference on Multimedia and Ubiquitous Engineering, Loutraki, Greece, June 28-30, 2011: 48-52.

[9] Bardeli R. Similarity search in animal sound databases [J]. IEEE Transactions on Multimedia, 2009, 11(1): 68-76.

[10] Huang Shaolin. The Design and Implementation of Indexing System Based on Lucene [J]. Journal of Modern Information, 2009, 24(28): 430-433.

[11] Bruce C, Donald M, and Trevor S. Search Engines: Information Retrieval in Practice [M]. Upper Saddle River: Addison-Wesley, 2010: 22-23.

[12] Han Ji-qing, Feng Tao, Zheng Gui-bin, et al. Audio Information Processing Technology [M]. Beijing: Tsinghua University Press, 2007: 32-46.

[13] Dan Zhao, Xiangdong Wang, Yueliang Qian. Fast Commercial Detection based on Audio Retrieval [C]/ Proceedings of the 2008 IEEE International Conference on Multimedia and Expo, ICME 2008, June 23-26 2008, Hannover, Germany. IEEE, 2008.

[14] C. Pedraza, J. Vitola, Johanna Sepulveda. Fast content-based audio retrieval algorithm [C]// Image, Signal Processing, and Artificial Vision (STSIVA), 2013 XVIII Symposium of. IEEE, 2013.