CLUSTERING OF IMPERFECT TRANSCRIPTS USING A NOVEL SIMILARITY MEASURE

Oktay Ibrahimov[1], Ishwar Sethi[1] and Nevenka Dimitrova[2]

[1]Intelligent Information Engineering Laboratory
Department of Computer Science & Engineering
Oakland University, Rochester, MI 48309
{ibrahimo, isethi}@oakland.edu

[2]Philips Research
345 Scarborough Road
Briarcliff Manor, NY10510-2099
nevenka.dimitrova@philips.com

ABSTRACT

There has been a surge of interest in last several years in methods for automatic generation of content indices for multimedia documents, particularly with respect to video and audio documents. As a result, there is much interest in methods for analyzing transcribed documents from audio and video broadcasts and telephone conversations and messages. The present paper deals with such an analysis by presenting a clustering technique to partition a set of transcribed documents into different meaningful topics. Our method determines the intersection between matching transcripts, evaluates the information contribution by each transcript, assesses the information closeness of overlapping words and calculates similarity based on Chi-square method. The main novelty of our method lies in the proposed similarity measure that is designed to withstand the imperfections of transcribed documents. Preliminary experimental results using an archive of transcribed news broadcasts demonstrate the efficacy of the proposed methodology.

1. INTRODUCTION

The field of multimedia information retrieval has seen a phenomenal growth in the last decade. The need for methods that can automatically generate content indices for video and audio documents has brought together researchers from many disciplines including image and speech processing, machine learning and pattern recognition, natural language processing, and information retrieval. The early efforts in this field were focused mainly on analyzing images and videos. These efforts have led to a set of powerful methods for video segmentation, key-frame extraction, camera motion characterization etc. However, the experience has been that no single modality of a multimedia document alone can yield rich enough content indices to build multimedia information retrieval systems that can satisfy the needs of a broad range of users. Consequently many researchers have explored the use of closed captions and audio to build more useful and reliable content indices.

While the methods analyzing closed caption text are closely tied to video analysis, the situation in the case of audio is different. Since an audio document can exist on its own, for example the recording of a radio broadcast, or as an integral part of a multimedia document, for example the soundtrack of a video or TV news broadcast, the interest in methods for analyzing audio documents have consequently focused on two main directions. First, there are researchers who have looked at soundtracks to build indices that can complement information extracted from the picture-track of a video. Examples of work in this category include several general audio data classification schemes that have been proposed to segment an audio document into coherent chunks of different types of audio classes – music, speech, speech and music etc. The second group of researchers has been more interested in generating transcripts of audio documents through automatic speech recognition and their analysis for automatic indexing. An example of work of this type is the work by Coden and Brown [1] who have been investigating the use of IBM ViaVoice speech recognition software by building a special acoustic model for broadcast news. The efforts in this category are many more. These have been mainly reported at TREC (Text Retrieval Conference organized annually by National Institute of Standards and Technology) under the label of Spoken Document Retrieval (SDR). These efforts have shown the possibility of applying automatic speech recognition technology to audio broadcasts to perform indexing and retrieval with a good degree of success. However, the success of many of these methods depends upon the size and quality of transcription.

Our interest in the area of audio analysis spans both of the above stated directions. We are interested in analyzing audio information to supplement video information. An example of such a recent work is the person identification in TV programs work of Li et al [2] of our group where speaker identification and face detection and recognition techniques are used in a complementary manner. We are also interested in making use of automatic speech recognition technology to obtain transcripts for audio chunks that are labeled as speech chunks for further analysis and indexing. Since such transcribed chunks are expected to exhibit transcription errors, poor topic boundaries, and small size, we have been investigating methods for suitably measuring similarity between such chunks of transcribed documents. In this paper we present the initial results of such an investigation using a novel similarity measure based on Chi-Square test. We use the proposed measure in a clustering algorithm to segment an archive of audio news broadcasts. Our results show that the suggested similarity measure is robust enough to be used for poorly transcribed speech segment for multimedia indexing applications.

In section 2 we introduce a novel similarity measure between transcripts. In section 3 we discuss our document clustering algorithm. In section 4 we present the experimental results and in section 5 we conclude the paper.

2. MEASURING SIMILARITY BETWEEN TRANSCRIPTS

In general, automatically transcribed text suffers from several additional problems not present in traditional text retrieval systems. These include transcription errors, for example the occurrence of inappropriate, erroneous words, ill-defined boundaries between different topics, and additional ambiguity or loose use of expressions inherent in conversational speech in contrast with written text. The fact that the same amount of information could be delivered by different sets of words contributes additional difficulty and makes the problem more complicated.

Taking into account the above set of difficulties, our approach for measuring similarity consists of the following steps:

1. **Transcript intersection**: Determine the intersection through word co-occurrences between matching transcripts.
2. **Information contribution**: Evaluate the amount of information contributed by every matching document to the intersection.
3. **Informative closeness**: Assess informative closeness of overlapping words;
4. **Similarity measure**: Calculate the similarity of matching documents.

2.1 Transcript intersection

In our approach we make an assumption that the amount of information contained in a document could be evaluated via summing the amount of information contained in the member words. Similarly, the amount of information contained in some part of a document might be evaluated via summing the amount of information contained in the corresponding words. For words, we assume that the amount of information conveyed by a word can be represented by means of the weight assigned to it. There are a number of methods that have been developed for weighting of words [3, 4]. In our approach we use the Okapi technique [3], which has proved to be efficient in a number of applications [5,6]. Thus, we will calculate the *Combined Weight* of a word by formula (1):

$$CW(w_i \mid D_j) = \frac{(K+1) * CFW(w_i) * TF(w_i, D_j)}{K * ((1-b) + b * NDL(D_j)) + TF(w_i, D_j)} \quad (1)$$

where $CFW(w_i) = \log(\frac{N}{n(w_i)})$ is the collection frequency weight, N is the total number of documents and $n(w_i)$ is the number of documents containing the word w_i. The quantity $TF(w_i, D_j)$ is the frequency of word w_i in the document D_j and $NDL(D_j)$ is the length of the document D_j normalized by the mean document length. The empirically determined constant b controls the influence of document length and is equal to 0.75. Another constant K may be viewed as a discounting parameter on the word frequency: when K is 0, the combined weight reduces to the collection frequency weight; as K increases the combined weight asymptotically approaches *tf*itf* [3]. In our case K is equal to 2.

Now, in accordance with assumptions stated above, we can easily get the weight of a document and weights of any of its parts via applying the formula (2).

$$DW(D_i) = \sum_{w_k \in X_i} CW(w_k) \qquad (2)$$

To obtain co-occurring words between documents, we consider the fact that not all words in documents are equally informative. Further, we take into account the rather high probability for erroneous words found in automatically transcribed documents. Thus, we first sort all words in transcripts by their weights and retain only those whose weights are greater than some preset threshold (this threshold has been determined empirically). These words are the only words considered for co-occurrence. By doing this we make a tacit assumption that there is a little chance for erroneous words to appear in a text in systematic way and as a result they should get less weight and, in general, not appear in the top of the sorted words.

2.2 Information contribution

As the words appearing in the intersection of documents generally convey different amount of information with respect to the documents to which they belong, we estimate the amount of information conveyed by every document to the intersection (3):

$$INTER(D_i, D_j) = \frac{DW(D_i \cap D_j)}{DW(D_i)} \qquad (3)$$

It is easy to derive from (3) the following inequality, which will be generally true when $D_i \neq D_j$.

$$INTER(D_i, D_j) \neq INTER(D_j, D_i) \qquad (4)$$

2.3 Informative closeness

Having determined the common words, we next evaluate informative closeness of the words appearing in intersection. This is done by representing the matching documents via their histograms. To evaluate informative similarity of the words belonging to the intersection in respect to matching documents, we apply Chi-square technique in a slightly reverse way. To carry out this step, we use the assumption that words w_k of the document D_i with the corresponding weights $CW(w_k \mid w_k \in D_i)$ constitute the set of words with the expected distribution of weights, and, the same words $w_k^{'}$ but belonging to the document D_j with the weights $CW(w_k^{'} \mid w_k^{'} \in D_j)$ constitute the set of words with the observed distribution of weights. Finally, we assume that null hypothesis, stating that two sets fit each other with some value of significance, is true. Thus, we can determine the significance value making our hypothesis true.

Through calculating by formula (5) chi-square values for observed and expected sets of words, and matching the value of χ^2 with the critical values for χ^2 through the standard table, we can easily find a significance value δ that will make our null hypothesis true.

$$\chi^2 = \sum_{w_k \in D_i \cap D_j} \frac{(CW(w_k \mid w_k \in D_i) - CW(w_k \mid w_k \in D_j))^2}{CW(w_k \mid w_k \in D_j)} \qquad (5)$$

Now having all necessary components we can calculate the similarity between two matching documents applying the formula (6):

$$sim(D_i, D_j) = \delta * INTER(D_i, D_j) \qquad (6)$$

Obviously, for similarities (6) we have the following inequality, which will get the value 1 if and only if $D_i = D_j$:

$$0 \leq sim(D_i, D_j) \leq 1 \qquad (7)$$

3. TRANSCRIPT CLUSTERING

In order to develop content indices, one of our goals is to be able to identify topics in a stream of transcripts. To develop a database of topics in an unsupervised manner, we have explored the use of above similarity measure in a sequential clustering procedure with the notion of "informative field" of a document. By informative field we reflect the range of information provided by a document. A cluster is, thus, a set of documents with "similar informative fields". We consider two documents D_i and D_j to be informative similar if $sim(D_i, D_j) > \tau$, where τ is some threshold. In our case τ =0.15.

The basic idea of the algorithm consists in determining of centers of distinguishing informative fields – centroids – and then finding the related documents.

The centroids in our case are also documents with the condition that a document chosen as a centroid of a cluster occupies the most of the information field associated with the cluster. The main steps of the clustering algorithm are as follows:

1. Let *k=1* be the index of the current cluster under construction, and *i=1* be the index of the current document.

2. Suppose that document D_i is the centroid C_k^* of cluster *k*.

3. Determine all documents, which are similar to C_k^* using the similarity measure described in 2.4.

4. Among all documents D_j determined in the previous step, find the document which gives the lowest value for the following ratio:

$$Di^0 = \min_j \{ \frac{sim(D_j, C_k^*)}{sim(C_k^*, D_j)} \} \qquad (8)$$

   Let i^0 be the index of the document thus found.

5. If $i \neq i^0$, assign the value of i^0 to *i* and go to the step 2.
6. Otherwise, the centroid for the current cluster has been found. Mark all documents determined at the step 3 as members of the cluster *k* and increment *k*.
7. Find the first document that doesn't belong to any clusters determined earlier and set its index to *i*. Go to the step 2.

4. EXPERIMENTAL RESULTS

Two sets of experiments have been performed. The first experiment was performed to see how well the similarity measure works in presence of transcription errors. In this experiment, we compute similarity between an original document, read by an American native speaker, and two other transcribed versions. All transcripts were obtained by reading stories to an IBM ViaVoice 1998 system. Documents dealing with two types of stories, business and politics were used in the experiments. The word error rate for the transcribed documents was found to vary in the interval of 2-10%. Table 1 shows the values of the similarity measure obtained. The highest value of similarity is 0.995 because the significance value δ used in the chi-square test is 0.995. The first number in a cell shows the similarity between original document and current transcript whereas the second number shows the reverse similarity.

Table 1.

	Original Reader	Reader A	Reader B
Business1	0.995 / 0.995	0.6457 / 0.6360	0.6467 / 0.5856
Business2	0.995 / 0.995	0.7057 / 0.6546	0.6482 / 0.5584
Politics1	0.995 / 0.995	0.8622 / 0.7295	0.8577 / 0.8162
Politics2	0.995 / 0.995	0.7783 / 0.6514	0.7666 / 0.6838

The numbers in Table 1 indicate that the suggested measure is good at capturing the similarity between documents in spite of the transcription errors. To further show that the measure is good at discriminating between different types of documents, we show in Table 2 the pair-wise similarities between the original versions of the four documents of Table 1.

Table 2.

	Busin.1	Busin.2	Polit.1	Polit.2
Busin.1	0.9950	0.3402	0.0221	0.0076
Busin.2	0.2333	0.9950	0.0613	0.0328
Polit.1	0.0194	0.0778	0.9950	0.1881
Polit.2	0.0135	0.0493	0.1851	0.9950

The second experiment was performed to evaluate how well the similarity measure will work with the clustering procedure outlined earlier. This experiment was done using transcripts from TDT 2 Careful Transcription Text Corpus, LDC catalog number LDC2000T44 [7]. The TDT2 corpus was created to support three TDT2 tasks: find topically homogeneous sections (segmentation), detect the occurrence of new events (detection), and track the reoccurrence of old or new events (tracking). The corpus contains 540 transcriptions of the broadcast news from ABC and CNN from January through June 1998 and of Voice of America (VOA) news broadcasts from March through June 1998.

For our experiment we picked out an arbitrary subset of transcripts from the TDT 2 corpus. These were manually clustered. The set of selected documents contained 202 documents with an average length of 72 words per document. The total number of words in the set after stopping and table look-up (derived words are replaced with corresponding basic words) was over 17000. The collection of treated documents is available at http://ieelab-secs.secs.oakland.edu . Then we applied our algorithm to the same set of transcripts to get the clusters of documents automatically. The clustered thus obtained were compared with clusters obtained manually to determine the precision and recall values. Table 3 presents the results for one group of topics.

Table 3.

Topic	*Precision*	*Recall*
Iraq, UN sanctions	100%	100%
Iraq, M.Albright talks	100%	100%
US actions against Iraq	100%	100%
M.Lewinsky, B.Clinton	100%	85.8%
India, Nuclear bomb test	100%	100%
Stock Market reports	100%	100%
Weather reports, Storm	100%	100%

Although these results are encouraging, we need to perform experiments with a larger archive to see how effective is the suggested approach. In our current implementation, we use a simple vocabulary containing about 7K basic words (about 30K with derived words). We feel that a larger vocabulary would improve the performance since after substitution of derived words with corresponding basic ones the preprocessed documents will be cleaner and the final results more accurate.

5. SUMMARY

A method to measure similarity between documents has been presented in this work. The suggested similarity measure is designed to cope with imperfectly transcribed documents. A method for clustering of the documents was also presented and applied to a carefully transcribed corpus. The preliminary results suggest that the suggested similarity measure and the clustering method are capable of achieving high precision and recall rates. We are in the process of further evaluation of our suggested similarity measure and its application to multimedia document indexing and retrieval.

ACKNOWLEDGEMENTS

The support of Philips Research is gratefully acknowledged.

REFERENCES

1. Anni R. Coden, Eric W. Brown, Speech Transcript Analysis for Automatic Search, In *IBM Research Report, RC 21838 (98287*), September 2000.

2. Li, D., G. Wei, I.K. Sethi, and N. Dimitrova, Fusion of Visual and Audio Features for Person Identification in Real Video, In Proc. Of the SPIE/IS&T Conference on Storage and Retrieval for Media Databases, pp. 180-186, San Jose, California, January 2001.

3. S.E. Robertson, K. Sparck Jones, Simple, Proven Approaches to Text Retrieval, http://www.uky.edu/~gbenoit/637/SparckJones1.html

4. M.Singler, R.Jin, A.Hauptmann, CMU Spoken Document Retrieval in Trec-8: Analysis of the role of Term Frequency TF, In *The 8th Text REtrieval Conference*, NIST, Gaithersburg, MD, November 1999.

5. Abberley, D., Renals, S., Cook G., Retrieval of broadcast news documents with the THISL system, In *Proc. of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, pp. 3781-3784, 1998.

6. Johnson, S.E., Jourlin P., Moore G.L., K.Sparck Jones, Woodland P.C., The Cambridge University Spoken Document Retrieval System. In *Proc. of the IEEE International Conference on Acoustic, Speech, and Signal Processing*, pp. 49-52, 1999.

7. http://www.ldc.upenn.edu/Catalog/TDT.html