行政院國家科學委員會專題研究計畫　成果報告

中文口語處理技術之前瞻性研究課題(3/3)

計畫類別：個別型計畫
計畫編號：NSC93-2213-E-002-015-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：國立臺灣大學電信工程學研究所

計畫主持人：李琳山

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 94 年 9 月 20 日

行政院國家科學委員會專題研究計畫期中報告

中文口語處理技術之前瞻性研究課題 （ 3/3 ）

計畫編號：NSC 93-2213-E-002-015-
執行期限：93 年 8 月 1 日至 94 年 7 月 31 日
主持人：李琳山　國立台灣大學電信工程學研究所
E-mail: lslee@gate.sinica.edu.tw

ABSTRACT

The most attractive form of future network content will be multi-media including speech information, and such speech information usually carries the core concepts for the content. As a result, the spoken documents associated with the multi-media content very possibly can serve as the key for retrieval and browsing. This paper presents a new approach of hierarchical topic organization and visual presentation of spoken documents for such a purpose based on the Probabilistic Latent Semantic Analysis (PLSA). With this approach the spoken documents can be organized into a two-dimensional tree (or multi-layered map) of topic clusters, and the user can very efficiently retrieve or browse the network content or associated spoken documents. Different from the conventional document clustering approaches, with PLSA the relationships among the topic clusters and the appropriate terms as the topic labels can be very well derived. An initial prototype system with Chinese broadcast news as the example spoken documents including automatic generation of titles and summaries and retrieval/browsing functionalities is also presented. Choice of different units other than words to be used as the terms in the processing is also considered in the system based on the special structure of the Chinese language.

中文摘要

多媒體資訊是未來網路資訊中最吸引人的一部份。語音則在其中扮演了最核心的一個角色。伴隨著多媒體資料而產生的語音文件將非常有可能在瀏覽以及擷取上成為鍵。這篇論文在語音文件的主題整合與視覺呈現方面提出了一個架構在機率式潛藏語音分上新的方法。經由此過程，語音文件能夠被整合進一個自行組織的二維空間主題叢集，以利使用者有效率地瀏覽或擷取資訊。和以往方式不同地，以機率式潛藏語音分析所取得的資訊，能夠同時提供主題間的關係，同時對於每個主題，也能夠找出適合的字詞來做為代表。初始的原型系統是建立在中文廣播新聞之上，配合了自動產生的摘要以及標題。不同的特徵選擇也使得這個新的方法能夠更佳的處理語音辨識上產生的錯誤。

1. INTRODUCTION

Speech is the primary and the most convenient means of communication between people [1]. In the future network era, the digital content over the network will include all the information activities for human life, from real-time information to knowledge archives, from working environments to private services, etc. Apparently, the most attractive form of the network content will be in multi-media including speech information, and such speech information usually tells the subjects, topics and concepts of the multi-media content. As a result, the spoken documents associated with the network content will become the key for retrieval and browsing. However, unlike the written documents with well structured paragraphs and titles, the multi-media and spoken documents are both very difficult to retrieve or browse, since they are just audio/video signals and the user can not go through each of them from the beginning to the end during browsing. A possible approach then may be to generate automatically titles and/or summaries for the spoken documents, analyze and organize the topics and concepts described in the spoken documents into some hierarchical structure, and then present the spoken documents in some visual form convenient for efficient retrieval/browsing applications.

The purpose of topic analysis and organization for spoken documents is to offer an overall knowledge of the semantic content of the entire spoken document archive in some form of hierarchical structures with concise visual presentation. The purpose is to enable comprehensive and efficient access to the spoken document archive, and to help the users to browse across the spoken documents efficiently. BBN's Rough'n'Ready system [2] may represent one of the few earliest efforts in this direction. The WebSOM method [3, 4] is another typical example towards data-driven topic organization for documents. In this method, the documents are clustered based on the self-organizing map (SOM) approach, and the relationships among the clusters can be presented as a two-dimensioned map describing the relationships among the topic clusters. The ProbMap [5] is a different approach with similar purpose but based on the Probabilistic Latent Semantic Analysis (PLSA) framework [6], in which the documents are organized into latent topic clusters, and the relationships among the clusters is presented as a two-dimensional map. Probabilistic Latent Semantic Analysis (PLSA) is an efficient approach developed for information retrieval purposes [6], in which a set of "latent topic variables", $\{T_k, k = 1, 2, \ldots, K\}$, is introduced, and all terms and documents are related to these latent topic variables in some probabilistic form.

In this paper, we present a new approach to analyze and organize the topics of spoken documents in an archive into a hierarchical two-dimensional tree structure or a multi-layer map for efficient browsing and retrieval. The basic approach used here, referred to as the Topic Mixture Model (TMM) [7] in this paper, is based on the PLSA concept but with slightly different formulation. A prototype system with Chinese broadcast news taken as the example spoken documents is also presented with rigorous performance evaluation. Special structure of Chinese language is also considered in the system.

2. Topic Mixture Model

To analyze the topic information about spoken documents and cluster them accordingly, a Topic Mixture Model (TMM) was developed here in this paper based on the well known PLSA concept but with slightly different formulation. In this model, each individual spoken document D_i is modeled as a probabilistic mixture for the terms:

$$P(t_j|D_i) = \sum_{k=1}^{K} P(t_j|T_k)P(T_k|D_i), \quad (1)$$

where t_j is a term (a term is a word in most cases, but can be phrases or other sub-word units as well) , $\{T_k, k = 1, 2, \ldots, K\}$ is the set of K "latent topics" as in PLSA, $P(t_j|D_i)$ is the probability of observing a term t_j in a document D_i , $P(t_j|T_k)$ is the probability of observing the term t_j for a specific latent topic T_k , and $P(T_k|D_i)$ is the probability (or weight) for the topic T_k being addressed by the document D_i , with the constraint $\sum_{k=1}^{K} P(T_k|D_i) = 1$. We then define a set of probability distributions $\{P(T_l|T_k), k = 1, 2, \ldots, K\}$ which represents the statistical correlation between the latent topic T_l and each of the other latent topics T_k. These distributions not only describe the semantic similarity among the latent topics, but can blend-in additional semantic contributions from related latent topics T_k to a given latent topic T_l. These probability distributions $P(T_l|T_k)$ can be expressed as a neighborhood function in terms of the distance between the locations of the latent topic T_l and those for other latent topics T_k on the two-dimensional map,

$$P(T_l|T_k) = \frac{exp[-d(T_l, T_k)^2/2\sigma_r{}^2]}{\sum_{m=1}^{K} exp[-d(T_m, T_k)^2/2\sigma_r{}^2]}, \quad (2)$$

where

$$d(T_l, T_k) = \sqrt{(x_l - x_k)^2 + (y_l - y_k)^2}, \quad (3)$$

is simply the Euclidean distance between the locations of the two points for T_l and T_k on the map with coordinates (x_l, y_l) and (x_k, y_k) , and the value of σ_r decreases as the number of iterations r of the EM algorithm described below increases. In this way, the conditional probability of observing a term t_j in a document D_i, $P(t_j|D_i)$ previously expressed in equation (1), can be modified as:

$$P(t_j|D_i) = \sum_{k=1}^{K} P(T_k|D_i)[\sum_{l=1}^{K} P(t_j|T_l)P(T_l|T_k)]. \quad (4)$$

This model is then trained in an unsupervised way with EM algorithm by maximizing the total log-likelihood L_T of the spoken document archive in terms of the unigram $P(t_j|D_i)$:

$$L_T = \sum_{i=1}^{N} \sum_{j=1}^{N'} n(t_j, D_i) log P(t_j|D_i), \quad (5)$$

where N is the total number of documents in the archive, N' is the total number of the different terms observed in the document archive. $n(t_j, D_i)$ is the number of times a term t_j occurring in the document D_i. The two probabilities in equation (1) can then be estimated by the expressions below:

$$P(t_j|T_k) = \frac{\sum_{D_i \in C} n(t_j, D_i) P_Z(T_k|t_j, D_i)}{\sum_{D_i \in C} \sum_{t_s \in D_i} n(t_s, D_i) P_Z(T_k|t_s, D_i)}, \quad (6)$$

$$P(T_k|D_i) = \frac{\sum_{t_s \in D_i} n(t_s, D_i) P_Y(T_k|t_s, D_i)}{|D_i|}, \quad (7)$$

where C is the corpus of the document archive, $|D_i|$ is the total number of terms in the document D_i , and

$$P_Z(T_k|t_j, D_i) = \frac{P(t_j|T_k)\sum_{l=1}^{K} P(T_k|T_l)P(T_l|D_i)}{\sum_{m=1}^{K} P(t_j|T_m)\left[\sum_{l=1}^{K} P(T_k|T_l)P(T_l|D_i)\right]}, \quad (8)$$

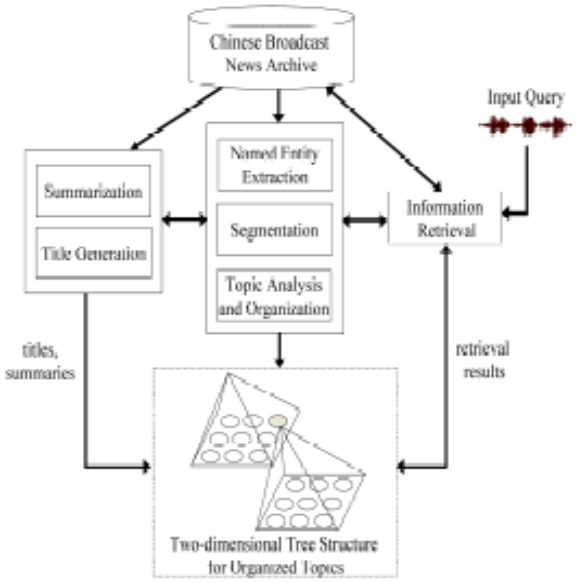

Figure 1: *Block diagram of the initial prototype system for Chinese broadcast news with efficient retrieval/browsing applications.*

$$P_Y(T_k|t_j, D_i) = \frac{P(T_k|D_i)\sum_{l=1}^{K} P(t_j|T_l)P(T_l|T_k)}{\sum_{m-1}^{K} P(T_m|D_i)\sum_{l=1}^{K} P(t_j|T_l)P(T_l|T_m)}. \quad (9)$$

There are at least two nice features of this approach as compared to conventional document clustering techniques. First, with the probability values the relationships among different topic clusters can be obtained and represented in the two-dimensional map. Second, the topic labels for each document cluster can be easily obtained by choosing the terms with the highest topic significance score $S(t_i, T_j)$ which is defined as the following,

$$S(t_i, T_j) = \frac{\sum_{k=1}^{N} n(t_i, D_k) P(T_j|D_k)}{\sum_{k=1}^{N} n(t_i, D_k)[1 - P(T_j|D_k)]}. \quad (10)$$

These nice features will be made clear in the following initial prototype system.

3. The Initial Prototype System for Chinese Broadcast News

Figure 1 is the block diagram of the initial prototype system for Chinese broadcast news. There are three parts in the system: the automatic generation of titles and summaries [8, 9, 10] is on the left, the information retrieval system is on the right, and in the middle is the topic analysis and organization presented in this paper, in which the news stories are well clustered into $m_1 \times m_1$ topics organized on a two-dimensional map, and each cluster of news stories can be further clustered into $m_2 \times m_2$ smaller topics with finer structure in the next layer and so on. This produces the two-dimensional tree structure for efficient retrieval/browsing.

The functionalities of the initial broadcast news retrieval and browsing prototype system are shown in Figure 2 and described below. First consider the top-down browsing functionalities. The homepage of the browsing system lists 20 news categories as in Figure 2(a) (not completely shown). When the user clicks the first category of "international political news", for example as shown here, a two-dimensional map of

3 *3 latent topic structure (with 9 blocks) appears as shown in Figure 2(b) (only 4 blocks are shown here), in which each block represents a major latent topic in the area of "international political news" for the news collection, characterized by roughly 4 topic labels shown in the block. As can be found, the block on the upper-right corner has labels "以色列(Israel)","阿拉法特(Arafat)", "巴勒斯坦(Palestine)" and "加薩市(Gaza City)", which clearly specify the topic. The block to its left, on the other hand, has labels "伊拉克(Iraq)", "巴格達(Baghdad)", "美軍(American Army)" and "陸戰隊(marine corps)", whereas the block below it in the middle-right has labels "聯合國(United Nations)", "安理會(Security Council)", "武檢人員(military inspectors)" and "武器(weapons)". Apparently, all these are different but related topics, and the distance between the blocks has to do with the relationships between the latent topics. The user can then click one of the blocks (for example, the one on the upper-right corner as shown here) to see the next layer 3*3 map for the fine structure of smaller latent topics for this cluster, as shown in Figure 2(c). As can be found in Figure 2(c), the block on the upper-right corner now has labels "以色列(Israel)", "夏隆(Shilom)", "約旦河(Jordan River)" and "美國(USA)", while the block below it has labels "中東(Middle East)", "鮑威爾(Powell)", "和平(peace)" and "路線(roadmap)", and so on. Apparently, the collection of broadcast news stories are now organized in a two-dimensional tree-structure or a multi-layer map for better retrieval and easier browsing. Here the second layer clusters are in fact the leaf nodes, therefore the user may wish to see all the news stories within such a node. With a click the automatically generated titles for all news stories clustered into that node are shown in a list, as in Figure 2(d) for the upper-middle small block in Figure 2(c) labeled with "阿拉法特(Arafat)" and so on, which includes the automatically generated titles for five news stories clustered into this block, together with the position of this node within the two-dimensional tree as shown in the lower-right corner of the screen. The user can further click the "summary" button after each title to listen to the automatically generated summaries, or click the title to listen to the complete news story. This two-dimensional tree structure with topic labels and the titles/summaries are therefore very helpful to browse the news stories.

The retrieval functionalities, on the other hand, are in general bottom-up. The screen of the retrieval system output for an input speech query (can be in either speech or text form), "請幫我找以色列與阿拉法特相關的新聞(Please find news stories relevant to Israel and Arafat)" is shown in Figure 2(e). As can be seen, a nice feature of this system is that all retrieved news stories, as listed in the upper half of Figure 2(e), have automatically generated titles and summaries. The user can therefore select the news stories by browsing through the titles, or listening to the summaries, rather than listening to the whole news story and then finding it was not the one he was looking for. The user can also select to click another functional button

Figure 2: *The top-down browsing and bottom-up retrieval functionalities of the initial prototype system.*

to see how a selected retrieved news item is located within the two-dimensional tree structure as mentioned previously in a bottom-up process. For example, if he selected and clicked the second item in the title list of Figure 2(e), "阿拉法特反對以色列所提結束包圍條件(Arafat objected to Israel's proposal for conditions of lifting the siege)", he can see the list of news titles in Figure 2(d) including the titles of all news stories clustered in that smaller latent topic (or leaf node), or go one layer up to see the structure of different smaller latent topics in Figure 2(c), or go one layer up further to see the structure of different major latent topics in Figure 2(b), and so on. This bottom-up process is very helpful for the user to identify the desired news stories or find the related news stories, even if they are not retrieved in the first step as shown in Figure 2(e).

4. Performance Evaluation

Very rigorous performance evaluation has been performed on the proposed approach based on the TDT-3 Chinese broadcast news corpus. A total of about 4,700 news stories in this corpus were used to train the model. A total of 47 different topics have been manually defined in TDT-3, and each news story was assigned to one of the topics manually, or assigned as "out of topic". These 47 classes of news stories with given topics were used as the reference for the evaluation presented below.

4.1. "Between-class to within-class" Distance Ratio

Intuitively, those news stories manually assigned to the same topic should be located on the map as close to each other as possible, while those manually assigned to different topics should be located on the map as far away as possible. We therefore define the "Between-class to within-class" distance ratio as in equation (11),

$$R = \bar{d}_B / \bar{d}_W \,, \tag{11}$$

where $\bar{d}_B$ is the average of the distance $d(T_l, T_k)$ in equation (3) over all pairs of news stories manually assigned to different topics, but located by the proposed algorithm to points (x_l, y_l) and (x_k, y_k) on the map here (thus is the "Between-class distance"), and $\bar{d}_W$ is the similar average, but over all pairs of news stories manually assigned to identical topics (thus the "Within-class distance"). So the ratio R in equation (11) tells how far away the news stories with different manually defined topics are separated on the map. Apparently, the higher values of R the better.

4.2. Total Entropy for Topic Distribution

For each news story D_i, the probability $P(T_k|D_i)$ for each latent topic $T_k, k = 1, 2, \ldots, K$, was given by the model. Thus the total entropy for topic distribution for the whole document collection with respect to the organized topic clusters can be defined as below:

$$H = \sum_{i=1}^{N} \sum_{k=1}^{K} P(T_k|D_i) \log(\frac{1}{P(T_k|D_i)}), \tag{12}$$

where N is the total number of news stories used in the evaluation. Apparently, lower total entropy means the news stories have probability distributions more focused on less topics.

Table 1: *Evaluation results for different choices as the "term" t_j.*

	Choice of Terms	Distance Ratio R	Total Entropy H
(a)	W	2.34	5135.62
(b)	S2	3.38	4637.71
(c)	C2	3.65	3489.21
(d)	S2+C2	3.78	4096.68

4.3. Test Results

Table 1 lists the results of the two performance measures proposed above. There are several choices of the "term" t_j used previously in section 2 considering the special structure of Chinese language, i.e., W(words), S2(segments of two syllables), C2(segments of two characters), and combinations. As we can see, the words (W in row(a)) were certainly NOT a good choice of terms for the purposes of topic analysis here. Segments of two syllables (S2 in row (b)) were apparently better with much higher distance ratio R and much lower total entropy H. Segments of two characters (C2 in row (c)) turned out to be even better. This is reasonable because in Chinese news many keywords useful in identifying the topics are new named entities or out-of-vocabulary (OOV) words, which very often can not be correctly recognized. On the other hand, in Chinese the syllables represent characters with meaning, and as a result in analyzing the topics of spoken documents the syllables make good sense even if not decoded into words which may not exist in the lexicon. In addition, each syllable may stand for many different homonym characters with different meanings, while a segment of two syllables very often gives very few, if not unique, polysyllabic words, and therefore the inherent topic. On the other hand, the one-to-many syllable-to-character mapping in Chinese implies that characters bring more precise information than syllables, if correctly decoded. These explained why S2 and C2 are better than words. The last row(d) indicated that integration of S2 and C2 may be another good choice, with better distance ratio R, though slightly higher total entropy H. In any case, when analyzing the Chinese spoken documents, segments of two syllables or two characters turn out to be more robust to recognition errors and provide better indication about the subject topic than words.

5. Conclusion

This paper presents a new approach for hierarchical topic analysis and organization for spoken documents, and the results are represented in a two-dimensional tree structure or a multi-layer map. This approach has been integrated with functionalities of automatic generation of titles and summaries and information retrieval to construct a single system for retrieving and browsing Chinese broadcast news. Very rigorous evaluation has been performed, and the results showed that when the special structure of Chinese language is considered, the approach and the system can be more robust to recognition errors, which is consistent with our previous work [10].

6. References

[1] B. H. Juang and S. Furui, "Automatic recognition and understanding of spoken language-a first step toward natural human-machine communication," Proceedings of the IEEE, vol. 88, no. 8, pp. 1142-1165, 2000.

[2] D.R.H. Miller, T. Leek and R. Schwartz, "Speech and language technologies for audio indexing and retrieval," Proc. IEEE, vol. 88, no. 8, pp. 1338-1353, 2000.

[3] T. Kohonen, S. Kaski, K. Lagus, J. Salojvi, J. Honkela, V. Paatero and Saarela A, "Self organization of a massive document collection," IEEE Trans on Neural Networks, vol. 11, no. 3, pp. 574-585, 2000.

[4] M. Kurimo, "Thematic indexing of spoken documents by using self-organizing maps," Speech Communication, vol. 38, pp. 29-45, 2002.

[5] T. Hofmann, "ProbMap - a probabilistic approach for mapping large document collections," Journal for Intelligent Data Analysis, vol. 4, pp. 149-164, 2000.

[6] Thomas Hofmann, "Probabilistic Latent Semantic Analysis," Uncertainty in Artificial Intelligence, 1999.

[7] B. Chen, Exploring the Use of Latent Topical Information for Statistical Chinese Spoken Document Retrieval ,'Minor revisions, Pattern Recognition Letters, January 2005.

[8] S. C. Chen and L. S. Lee, "Automatic title generation for Chinese spoken documents using an adaptive K-nearest-neighbor approach," in Proc. European Conference on Speech Communication and Technology, 2003, pp. 2813-2816.

[9] L. S. Lee and S. C. Chen, "Automatic title generation for Chinese spoken documents considering the special structure of the language," in Proc. European Conference on Speech Communication and Technology, 2003, pp. 2325-2328.

[10] L. S. Lee, Y. Ho, J. F. Chen and S. C. Chen, "Why is the special structure of the language important for Chinese spoken language processing? -examples on spoken document retrieval, segmentation and summarization," in Proc. European Conference on Speech Communication and Technology, 2003, pp. 49-52.