Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 27 (2006) 9–18

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Exploring the use of latent topical information for statistical Chinese spoken document retrieval

Berlin Chen [*]

Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University, No. 88, Section 4, Ting-Chow Road, Taipei 116, Taiwan, ROC

Received 1 October 2004; received in revised form 14 February 2005
Available online 29 August 2005

Communicated by E. Backer

Abstract

Information retrieval which aims to provide people with easy access to all kinds of information is now becoming more and more emphasized. However, most approaches to information retrieval are primarily based on literal term matching and operate in a deterministic manner. Thus their performance is often limited due to the problems of vocabulary mismatch and not able to be steadily improved through use. In order to overcome these drawbacks as well as to enhance the retrieval performance, in this paper, we explore the use of topical mixture model for statistical Chinese spoken document retrieval. Various kinds of model structures and learning approaches were extensively investigated. In addition, the retrieval capabilities were verified by comparison with the probabilistic latent semantic analysis model, vector space model and latent semantic indexing model, as well as our previously presented HMM/N-gram retrieval model. The experiments were performed on the TDT Chinese collections (TDT-2 and TDT-3). Noticeable improvements in retrieval performance were obtained.

Keywords: Information retrieval; Topical mixture model; Probabilistic latent semantic analysis model; Vector space model; Latent semantic indexing model; HMM/N-gram retrieval model

1. Introduction

Due to the advent of computer technology and the proliferation of Internet activity, tremendous volumes of multimedia information, such as text files, Web pages, broadcast radio and television programs, digital archives and so on, are continuously growing and filling our computers and lives. Development of intelligent and efficient retrieval techniques to provide people with easy access to all kinds of information is now becoming more and more emphasized (Voorhees and Harman, 2000). It is also obvious that speech is the primary and most convenient means of communication between people, as well as the most rich source of information for the great volumes of multimedia. Therefore, with the rapid evolution of speech recognition technology, substantial efforts and very encouraging results on recognition and retrieval of spoken documents have been reported in the last few years (Woodland, 2002; Gauvain et al., 2002; Beyerlein et al., 2002; Chen et al., 2002; Chang et al., 2002; Meng et al., 2004; Byrne et al., 2004).

The conventional information retrieval (IR) approaches in principle can be characterized from two major perspectives: the matching strategy and the learning capability. There are two matching strategies frequently used to determine the degree of relevance for a document with respect to a query, namely, literal term matching and concept matching. The vector space model (VSM) and probability-based model approaches are primarily based on literal term matching. VSM, which takes the vector representations of the query and documents, has been widely used because

[*] Tel.: +886 2 29322411x203; fax: +886 2 29322378.
E-mail address: berlin@csie.ntnu.edu.tw
URL: http://berlin.csie.ntnu.edu.tw

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.06.010

of its simplicity and satisfactory performance (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999). The probability-based approach instead attempts to handle the retrieval problem within a probabilistic framework. The language model (Ponte and Croft, 1998; Song and Croft, 1999; Zhai and Lafferty, 2001) and the hidden Markov model (HMM) (Miller et al., 1999; Lo et al., 2003; Chen et al., 2004a) are good examples of it, and research at a number of sites has confirmed that such a modeling approach does provide a potentially effective and theoretically attractive probabilistic framework for studying information retrieval problems. Excellent survey articles on the use of the probability-based approach for information retrieval can also be found (Croft and Lafferty, 2003; Liu and Croft, 2005). However, most of these approaches often suffer from the problem of word usage diversity (or so-called vocabulary mismatch), which will make the retrieval performance degrade severely as a given query and its relevant documents are using quite a different set of words. In contrast, concept matching is based on discovering the latent topical information embedded in the query and documents, and latent semantic indexing (LSI) model is one example. LSI transforms the high-dimensional vector representations of the query and documents into a lower dimensional space (the so-called latent semantic space). Then the similarity measure can be estimated in the reduced space, where a query and a document may have a high proximity value even if they do not share any words or terms in common (Furnas et al., 1988; Deerwester et al., 1990). On the other hand, from the perspective of learning capability, it is well known that VSM and LSI are based on linear algebra operations and can incorporate a wide range of term weighting schemes as well as query or document expansion formulae (Salton and Buckley, 1988; Sparck Jones et al., 1998; Singhal and Pereira, 1999; Mandala et al., 2000) to modify the representations of query or documents, or to improve the information retrieval performance. While the probability-based approach, such as HMM, follows solid statistical foundations for automatic model refinement or optimization (Makhoul et al., 2000; Liu and Croft, 2005; Chen et al., 2004a), and thus can be steadily improved by using a variety of machine learning algorithms in either supervised or unsupervised modes.

Based on these observations, in this paper we study the use of topical mixture model for statistical Chinese spoken document retrieval, which in essence belongs to the probability-based approach and has virtue of being able to perform concept matching as well. Various kinds of model complexities for the topical mixture model were extensively investigated. In addition, their retrieval capabilities were verified by comparison with the other retrieval models. Structures similar to the presented approach also have been investigated in the machine learning literature recently (Hofmann, 2001; Blei et al., 2003; Wang et al., 2005). There are several differences between the presented approach and the previous ones. First, we explicitly interpret the document as a mixture model used to predict the query, which can be easily related to the conventional HMM modeling approaches that have been widely studied in speech and language processing community, and quite a few of theoretically attractive model training algorithms or optimization criteria can be therefore applied (Chou and Juang, 2003, Chapter 1). Moreover, we measure the relevance between the query and documents directly under the likelihood criterion (or in the likelihood space), unlike the previous approach (Hofmann, 2001), which evaluates the relevance between the query and documents in the low-dimensional factor (topic) space and only reports results by linearly combining with the cosine measure score obtained by using the VSM retrieval model. Finally, in this paper, both the supervised and unsupervised model learning approaches are extensively studied, while in the previous work (Hofmann, 2001; Blei et al., 2003; Wang et al., 2005), only unsupervised learning was investigated. We find that the results obtained based on the supervised learning approach are much better than those based on the unsupervised one.

In this paper, all the experiments were performed on the Topic Detection and Tracking Corpora (TDT-2 and TDT-3). The TDT corpora have been used for cross-language spoken document retrieval (CL-SDR) in the Mandarin English Information (MEI) Project (Meng et al., 2004), which is an NSF sponsored project conducted at the Johns Hopkins University Summer Workshop 2000. Project MEI investigated the use of an entire English newswire story (text) as a query to retrieve relevant Mandarin Chinese radio broadcast news stories (audio) in the document collection. In this paper, we study the monolingual spoken document retrieval task instead. All the experiments were tested on the task involving the use of an entire Chinese newswire story (text) as a query to retrieve relevant Mandarin Chinese radio broadcast news stories (audio) from the document collection. Such a retrieval context is termed query-by-example. This technique can help users to find the corresponding video or audio news reports, which could be more attractive and informative when they see a newswire text report. Most of the prior work on Chinese spoken document retrieval is focused on retrieving spoken documents by short queries (Wang, 2000; Bai et al., 2001; Chang et al., 2002).

The rest of this paper is organized as follows. The experimental corpus is introduced in Section 2. In Section 3, we explain the structural characteristics of the topical mixture model and briefly review the other retrieval models. Then, the experimental settings and a series of information retrieval experiments are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Experimental corpus

We used two Topic Detection and Tracking (TDT) collections (LDC, 2000) for this work. TDT is a DARPA-sponsored program where participating sites tackle tasks such as identifying the first time a news story is reported