Pattern-Based Identification in P2P Systems

Gábor Richly
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

Ferenc Kovács
Budapest University of Technology and Economics, Hungary

INTRODUCTION

This article presents a novel approach to search in shared audio file storages such as P2P based systems. The proposed method is based on the recognition of specific patterns in the audio contents in such a way extending the searching possibility from the description based model to the content based model. The importance of the real-time pattern recognition algorithms that are used on audio data for content-based searching in streaming media is rapidly growing (Liu, Wang, & Chen, 1998). The main problem of such algorithms is the optimal selection of the reference patterns (*soundprints*) used in the recognition procedure. The proposed method is based on distance maximization and is able to quickly choose the pattern that later will be used as reference by the pattern recognition algorithms (Richly, Kozma, Kovács, & Hosszú, 2001). The presented method called *EMESE (experimental media-stream recognizer)* is an important part of a lightweight content-searching method, which is suitable for the investigation of the networkwide shared file storages. The experimental measurement data shown in the article demonstrate the efficiency of the proposed procedure.

BACKGROUND

From the development of *Napster* (Parker, 2004), the Internet based communication has been developed toward the application level networks (ALNs). On the more and more powerful hosts, various collaborative applications run and create virtual (logical) connections with each other (Hosszú, 2005). They establish virtual overlay, and as an alternative of the older client/server model, they use peer-to-peer (P2P) communication. The majority of such systems deal with file sharing; that is why their important task is to search in large, distributed shared file storages (Cohen, 2003; Qiu & Srikant, 2004).

Until now, the search has been usually carried out based on the various attributes of the media contents (Yang & Garcia-Molina, 2002). These metadata are the name of the media file, the name of the authors, data of recording, type of the media content, and maybe some keywords and other descriptive attributes. However, if the incorrect metadata were accidentally recorded, the media file may become invisible due to the misleading descriptions. Currently, the powerful computers give the possibility to implement and widely use pattern recognition methods. Naturally, due to the large amount of media files and their very rich content, very limited pattern identification should be reached as a realistic goal. This article introduces the problem of the media identification based on well-defined pattern recognition.

Another problem is introduced if the pattern-based identification method should be extended form media files to real-time media streams. The hardness of this problem is the requirement that the pattern identification system must work in real-time even in weak computing environments. For this purpose, the full-featured media monitoring methods are not applicable since they require the large processing power in order to run their full-featured pattern recognition algorithms.

The novel system named *EMESE* is dedicated for solving the special problem, where a small but significant pattern should be found in a large voice stream or bulk voice data file in order to identify known sections of audio. Since we limit our review to the sound files, the pattern, which serves for identifying the media content of a file, is named *soundprint*. The developed method is lightweight, meaning that its design goals were the fast operation and relatively small computing power. In order to reach these goals, the length of the pattern to be recognized should be very limited and the total score is not required. This article deals mainly with the heart of the EMESE system, the pattern recognition algorithm, especially with the creation of the reference pattern, the process called *reference selection*.

THE PROBLEM OF PATTERN RECOGNITION

In the field of sound recognition, there are many different methods and applications for specific tasks (Coen, 1995; Kondoz, 1994). The demand for working efficiently with streaming media on the Internet increases rapidly. These audio streams may contain artificial sound effects besides the mix of music and human speech. These effects furthermore may contain signal fragments that are not audible by the ear. As a consequence, processing of this kind of *audio signal* is rather different from the already developed methods, as for example the short-term predictability of the signal is not applicable. The representation of digital audio signal as individual sample values lacks any semantic structure to help automatic identification. For this reason, the audio signal is transformed into several different orthogonal or quasi-orthogonal bases that enable detecting certain properties.

Already, there are solutions for classifying the type of broadcast on radio or television using the audio signal. The solution in Akihito, Hamada, and Tonomura (1998) basically makes a speech/music decision by examining the spectrum for harmonic content and the temporal behavior of the spectral-peek distribution. Although it was applied successfully to that decision problem, it cannot be used for generic recognition purposes. Liu et al. (1998) also describe a scheme classifying a method where the extracted features are based on the short-time spectral distribution represented by a bandwidth and a central frequency value. Several other features, for example, the volume distribution and the pitch contour along the sound clip, are also calculated. The main difficulty of these methods is their high computation-time demand. That is why their application for real-time or fast monitoring is hardly possible when taking the great number of references to be monitored into account.

A similar monitoring problem was introduced in Lourens (1990) and the used feature, a section of the energy envelope of the record signal (*reference*), was correlated with the input (*test*) signal. The demand on real-time execution drove the development of the recognition scheme introduced in Richly, Varga, Hosszú, and Kovács (2000) that is capable of recognizing a pattern of transformed audio signal in an input stream, even in the presence of level-limited noise. This algorithm first selects a short segment of the signal from each record in the set of records to be monitored (Richly et al., 2001).

Carrying out tests on live audio broadcasts showed that the success of identification process depends on the proper selection of the representative short segment. The position where this representative segment can be extracted is determined by the recognition algorithm of the proposed system called EMESE. The selected references must be noncorrelated to avoid false alarms. The method applied in the novel method EMESE is analyzed in the following in order to check how the monitoring system can synchronize to the stream under various conditions, and the measured results are also presented.

THE SOUND IDENTIFICATION IN THE EMESE

The reference selection algorithm needs a well understanding of the recognition method. The audio signal, sampled at $f_s=16kHz$, is transformed into a spectral description. It is a block of data, where the columns are feature vectors of the sound corresponding to a *frame* of time-domain data ($N_f=256$ samples, $T_f=16ms$ long). First, the amplitude of the Fourier spectrum is computed from the frame. Then, averaging is adapted to the neighboring frequency lines to project the spectrum onto the Bark-scale. The reason for this is to speed up the later comparison stage and to include a well established emphasizing tool used in audio processing, the perceptual modeling of the human auditory system. As a result, we get $N_B=20$ values, building up a vector, that are normalized and quantized. Two levels are determined in each transformed frame. The levels are the 10% and 70% of the peak value of the amplitude spectrum. We name the transformed frame a *slice*. In every reference, there are N_S=50 slices of non-overlapping consecutive frames and the audio section, from

Related Content