Multimedia Summarization in Law Courts: A Clustering-Based Environment for Browsing and Consulting Judicial Folders

E. Fersini[1], E. Messina[1], and F. Archetti[1,2]

[1] DISCo, Università degli Studi di Milano-Bicocca,
Viale Sarca, 336 - 20126 Milano, Italy
{fersini,messina,archetti}@disco.unimib.it
[2] Consorzio Milano Ricerche,
Via Cicognara 7 - 20129 Milano, Italy
archetti@milanoricerche.it

Abstract. Digital videos represent a fundamental informative source of those events that occur during a penal proceedings, which thanks to the technologies available nowadays, can be stored, organized and retrieved in short time and with low cost. However, considering the dimension that a video source can assume during a trial recording, several requirements have been pointed out by judicial actors: fast navigation of the stream, efficient access to data inside and effective representation of relevant contents. One of the possible solutions to these requirements is represented by multimedia summarization aimed at deriving a synthetic representation of audio/video contents, characterized by a limited loss of meaningful information. In this paper a multimedia summarization environment is proposed for defining a storyboard for proceedings celebrated into courtrooms.

1 Introduction and Motivation

Multimedia summarization techniques analyze several informative sources comprises into a multimedia document, with the aim of extracting a semantic abstract. Multimedia summarization techniques available in literature can be divided in three main categories: (1) internal techniques, which exploit low level features of audio, video and text; (2) external techniques, which refer to the information typically associated with a viewing activity and interaction with the user; (3) hybrid techniques, which combine internal and external information.

These techniques are focused on different types of features: (a) domain specific, i.e. typical characteristics of a given domain known a priori and (b) non-domain specific, i.e. non-generic features associated with a particular context.

With respect to internal techniques the main goal is to analyze low-level features derived from text, images and audio contents within a multimedia document. Interesting example can be found in [1], [2] and [3]. In [1] the semantics of objects and events occurring within news video are extracted from subtitles and used to specialize / improve the systems of automatic speech recognition. In

P. Perner (Ed.): ICDM 2010, LNAI 6171, pp. 237–247, 2010.

[2] the performance related to the identification of special events are increased by combining scene recognition techniques with OCR-based approaches for subtitles recognition in baseball video documents. In [2] the scenes containing text in football videos are recognized using OCR techniques, for then a subsequent identification of key events through audio and video features.

In order to reduce the semantic gap between low level features and semantic concepts, research is moving towards the inclusion of external information that usually comprise knowledge about user-based information and the context in which a multimedia document evolves. The techniques able to generate a video summary on the basis of external information are limited to three case studies [4] [5] [6] focused on using domain specific features. In [4] a summarization technique is proposed in order to gather context information from the acquisition/registration phase, in particular by monitoring the movement of citizens around their houses. Cameras at a specific position and pressure sensors are used to track users. Since users are not required to provide any kind of information, the summary is produced by analyzing data concerning the movement (such as the distance between steps and direction changes). In [5] and [6] semantic annotations, collected during the production phase of the video and described by the standard MPEG-7, are analyzed. In particular in [6] a sequence of audio-video segments is produced on the basis of annotations from video sports (baseball matches), such as players' names or specific events occurring during the match. In [5] a video, characterized by a set of MPEG-7 macro-semantic annotations collected during the acquisition phase, is further annotated by users in order to indicate their level of interest in each video segment. The associations between preferences and the macro-annotations are then modelled by using supervised learning approaches to enable the generation of automatic summary of new multimedia documents.

An attempt that tries to combine the peculiarities of the previous techniques is represented by Hybrid Techniques. Hybrid summarisation techniques combine the advantages provided by internal and external approaches by analyzing a combination of internal and external information. As overviewed for the previous techniques, the hybrid ones can be distinguished in domain specific and non-domain specific. Examples of domain-specific hybrid techniques are related to music videos [7], broadcast news [8] and movies [9]. In non-domain specific approaches we can find two main investigations:

- in [10] the summarization approach could be described by two stages: (1) frames are grouped by a clustering approach, using colour image features; (2) during the editing phase, manual annotations of the representative frame of each cluster, with a subsequent spread to frames of the same cluster, are required. The summaryof is then generated by choosing those representative elements of each cluster matching the user query.
- in [11] an annotation tool is used during the editing phase in order to propagate semantic descriptors to non-labelled contents. During the summary generation phase, the user profile is considered in order to create a customized synthetic representation.

According to the output that a multimedia summarization technique should generate, we can distinguish between static and dynamic summaries. A static summary, also known also as storyboard, can be viewed as a series of key frames or video segments. Approaches for static summaries are focused on identifying relevant contents, do not considering the sequential aspect. They are described by a sub-sampling activity tuned according to the number of desired key-frames. Their use is related to hypermedia documents in order to access the internal parts of a multimedia source. This kind of summary should respect the following requirements: (1) conciseness, they do not to exceed a given limit related to the number of images (key frames), (2) content coverage, they should maximize (minimize) the similarity (dissimilarity) between images for the selection of key frames. A dynamic summary, also known as video skim, consists of a sequence of images associated to their soundtrack. They are generally presented as a video clip or trailer and can be viewed as a preview of the original video, where unimportant shots and scenes are omitted. This kind of summary should respect the following requirements: (1) conciseness, they do not to exceed a given time limit, (2) content coverage, they should maximize the temporal distribution of original video, (3) visual consistency, must minimize the frequency of changes of scene.

There are many differences between static and dynamic summary. The static video summary can be obtained more quickly than the dynamic one because it is focused on the use of only visual features, derived from the images that compose the video itself, without taking into account information from the audio stream. Consequently, once identified the key frames, the creation of the storyboard is a simple activity: audio/video synchronization is not required. A further advantage provided by the static summary is related to the temporal order of the frames: the user is able to quickly understand the contents of a video by looking directly at the sequence of the selected frames. Concerning with dynamic video summary, there are other types of benefits. Dynamic summaries, compared to static ones, use the information coming from the audio stream in a rational way: if on one hand there is a high computational complexity, on the other hand there is a gain in terms of meaning provided by the audio stream.

By analyzing the state of the art related to multimedia summarization techniques, no evidences about summaries over courtroom proceedings are given. The main reasons behind this lack are related to the characteristic of the judicial domain: (1) courtroom recordings are usually charachterized by low quality of audio and video sources; (2) significant events occurring during a debate are not characterized by low level features and therefore we need to understand semantic concepts of interest; (3) a very high level of compression is expected, implying a summary with 2-5 keyframes (which is difficult to derive only from images). For this reasons a comprehensive approach for tackling the current constraints need to be defined. In this paper we are mainly addressing the problem of deriving a storyboard of a multimedia document coming from penal proceedings recordings, by proposing an external summarization technique based on

the unsupervised clustering algorithm named Induced Bisecting K-Means. The main outline of this paper is the following. In section 2 the proposed multimedia summarization environment is presented. In section 3 the workflow for deriving a storyboard for the judicial actors is described. In section 4 details about the exploited clustering algorithm are given. Finally, in section 5 conclusions are derived.

2 Multimedia Summarization Environment

In order to address the problem of defining a short and meaningful representation of a debate that is celebrated within a law courtroom, we propose a multimedia summarization environment based on unsupervised learning. The main goal is to create a storyboard of either a hearing or an entire proceedings, by taking into account the semantic information embedded into a courtroom recording. In particular, the main information sources exploited for producing a multimedia summary are represented by:

- automatic speech transcriptions that correspond to what is uttered by the actors involved into hearings/proceedings. The automatic transcription are provided by Automatic Speech Recognition (ASR) systems, investigated in [14] [15], trained on real judicial data coming from courtrooms. Since it is impossible to derive a deterministic formula able to create a link between the acoustic signal of an utterance and the related sequence of associated words, the ASR system exploits a statistical-probabilistic formulations based on Hidden Markov Models [17]. In particular, a combination of two probabilistic models is used: an acoustic model able to represent phonetics, pronounce variability, time dynamics (co-utterance), and a language model able to represent the knowledge about word sequences.
- automatic audio annotations coming from emotional states recognition (for example fear, neutral, anger). The emotional state annotations are derived through a framework based on a Multi-layer Support Vector Machine approach [18]. Given a set of sentences uttered by different speakers, a features extraction step is firstly performed in order to map the vocal signals into descriptive attributes (prosodic features, formant frequencies, energy, Mel Frequency Cepstral Coefficients, etc...). These features are then used to create a classification model able to infer emotional states of unlabelled speakers.
- automatic video annotations that correspond to what happen during a debate (for instance change of witness posture, new witness, behavior of a given actor). The motion analysis of judicial videos is based on a combinations of video processing algorithms, in order to achieve reliable localization and tracking of significant features. In order to analyze the motions taking place in a video, and to track gestures or head movements of given subjects (typically the witnesses), the optical flow is extracted as the moving points. Then active pixels are separated from the static ones using a kurtosis-based method and finally through a wavelet based approach extracting relevant

features. At this stage the link between low level features and a given set of relevant actions is performed through the induction of Bayesian learner.

The Multimedia Summarization Environment includes two different modules: the acquisition module and the summarization module.

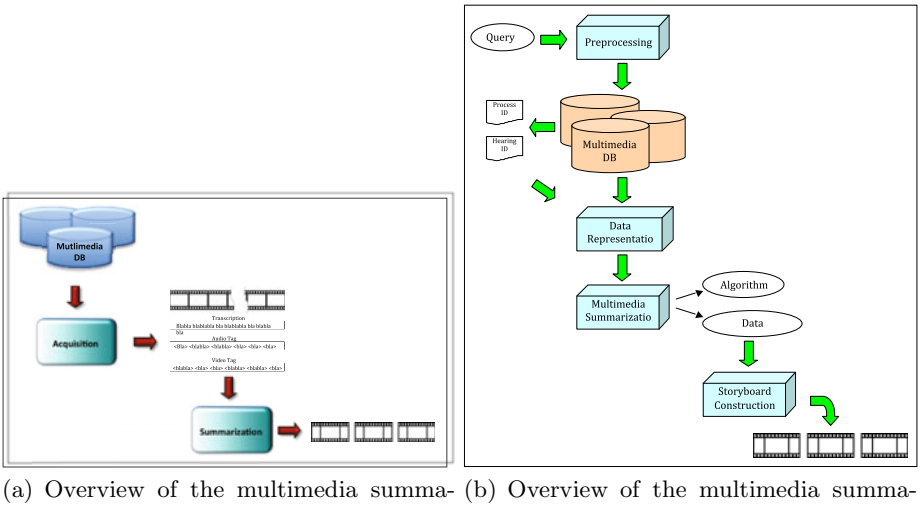
- The *acquisition module*, given a textual query specified by the end user, retrieves multimedia information from the Multimedia Database in terms of audio-video track(s), speech transcription and semantic annotations.
- The *summarization module* is aimed at producing a storyboard by exploiting the information retrieved by the acquisition module. The summary is created by focusing on maximally query-relevant passages and reducing cross-document redundancy.

A simple overview of the modules involved into the multimedia summarization environment is depicted in figure 1 (a).

3 Multimedia Summarization Workflow

In order to summarize a multimedia document according to the user needs, a query statement is specified to start the entire workflow (see figure 1 (b)).

The user query is specified at the graphic interface level, where a list of trials are available, in terms of keywords in which we are interested (whatever is uttered by the involved speaker, the emotional state of actors, etc...).



(a) Overview of the multimedia summarization macro-modules

(b) Overview of the multimedia summarization workflow

Fig. 1. Multimedia Summarizaion Environment

Once the query has been specified, it is submitted to the pre-processing module. The aim of this module is to optimize the user query by eliminating noise and by reducing the size of vocabulary, i.e. stop words removal and stemming are performed to enhance retrieval performance on transcription and annotations.

After the preprocessing activity the query is submitted to the retrieval module, which is aimed at accessing to the multimedia database, in order to identify all the information matching the user query: transcription of the debate, audio annotations and videos annotations. At this level, two possibilities are given to the end user: to summarize an entire trial or only those sub-parts of the proceedings that match the query. In the first case the user query is used to retrieve the multimedia documents related to a trial by executing a high-level skimming of the overall database. After this initial step all the clips of the retrieved hearing are considered for producing the summary. In the second case the query is used to scan the database in a more exhaustive way so that, within a given trial, only the audio, video and textual clips that completely match the user query are retrieved.

In both cases we refer to a (audio, video and textual) clip as a consecutive portion of a debate in which there is one speaker whom is active, i.e. there exist a sequence of words uttered by the same speaker without breaking due to other speakers. Indeed, a clip compreses a textual transcription for each speaker period with the corresponding audio/video tags.

The next step in the multimedia summarization workflow relates to data representation module. The aim of this module is to combine information coming from different sources in order to create a unified representation. This activity is performed through a feature vector representation, where all the information able to characterize the audio, video and textual clip of interest are managed as features and weights. Examples of features exploited by this representation are given by the textual transcription, the audio and video tag, the start and end time of the relevant sub-parts of the debate. In particular, two matrices are defined to be exploited by the summarization module: one matrix

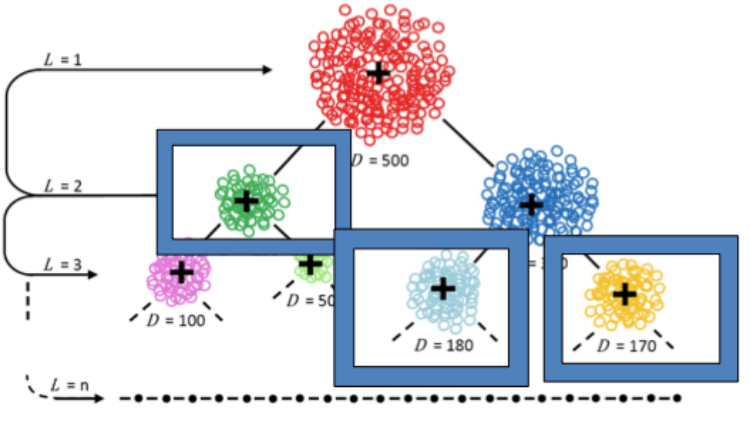


Fig. 2. Dichotimic tree generated by Induced Bisecting K-Means

associated with speech transcription and one matrix associated to the audio and video annotations. The first matrix, defined as numerical, represents textual transcription scoring, obtained through the TFIDF weighting technique [19]. A speech transcription segment associated to a single speaker is mapped into a row, while each term is mapped into a column. The second matrix, defined as binary, represents the presence or absence of a specific audio/video annotation associated to a transcription.

Starting from these two matrices, the multimedia summarization module may start the summary generation. The core component is based on a clustering algorithm named Induced Bisecting K-means [13]. The algorithm creates a hierarchical organization of (audio, video and textual) clips, by grouping in several clusters hearings (or sub-parts of them) according to a given similarity metric. This algorithm is able to build a dichotomic tree in which coherent concepts are grouped together, i.e. each cluster created by the algorithm contains a set of audio, video and textual clips representing similar concepts that are coherent with the user query (see figure 2).

The last step relates to the storyboard construction, where the final storyboard is derived from the dichotomic tree structure produced by the Induced Bisecting K-means algorithm. Given the dichotomic tree, a pruning step is performed in order to choose only those clusters that satisfy a given intra-cluster similarity requirements [20]. Suppose that the pruning activity after the Induced Bisecting K-means returns a set of clusters as reported in figure 3 where C1, C2 and C3 are the resulting clusters and the clips named $1, \ldots, 9$ represent the sub-parts of the debate. The storyboard construction activity considers the representative elements of each cluster (centroids) as the relevant clips for the summary. The storyboard is generated by presenting to the end user the first frame of each centroid, connected to the corresponding audio, video and textual

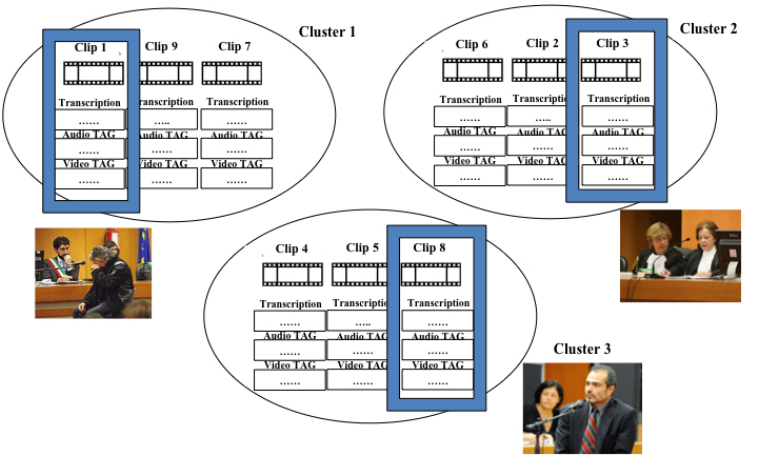


Fig. 3. Clustering output

information, references of the trial/hearing, start and end time of the segments and so on. By referencing figure 3, only the first frames related to segments 1, 3 and 8 (representative of the obtained 3 clusters) are presented to the end user as pictures that could be clicked to start the corresponding audio-video portion.

In the following subsection details about the core component of the multimedia summarization environment, i.e. the Induced Bisecting K-Means clustering algorithm, are given.

4 The Hierarchical Clustering Algorithm

The approaches proposed in the literature for hierarchical clustering were mostly statistical with a high computational complexity . A novel approach, Bisecting k-Means was proposed in [12], has a linear complexity and is relatively efficient and scalable. It starts with a single cluster of multimedia clips and works in the following way:

Algorithm 1. Bisecting K-Means

1: Pick a cluster S of clips m_l to split
2: Set K as the number of clusters to be obtained
3: Select two random seeds which are the initial representative clips (centroids)
4: Find 2 sub-clusters S1 and S2 using the basic k-Means algorithm[1].
5: Repeat step 2 and 3 for ITER times and take the split that produces the clustering with the highest Intra Cluster Similarity (ICS)[1]
6: $ICS(S_p) = \frac{1}{|S_p|^2} \sum_{m_i, m_j \in S_p} sim(m_i, m_j)$
7: Repeat steps 1, 2 and 3 until the desired number of clusters is obtained.

The major disadvantage of this algorithm is related to the requirements about the specification (a priori) of the parameters K and ITER. An incorrect estimation of K and ITER may lead to poor clustering accuracy. Moreover, the algorithm is sensitive to the noise tath may affect the computation of cluster centroids. Consider for instance N as the number of clips belonging to the cluster p and R as the set of their indices. The j^{th} feature of the cluster centroid - used by the k-Means algorithm during step 3 - is computed as $c_j^p = \frac{1}{N} \sum_{r \in R} m_{rj}$ where m_{rj} is the vectorial representation of the j^{th} feature of the i^{th} clip. Consequently, the centroid c_j^p may contain the contribution of noisy features that the pre-processing phase is not be able to remove. To overcome these two problems we exploit an extended version of the Standard Bisecting k-Means, named Induced Bisecting k-Means [13], whose main steps are described as follows:

[1] The similarity metric is a linear combination of the cosine similarity, for the numerical vectors concerned with transcriptions, and the jaccard similarity, for the binary vectors concerned with audio/video annotations.

Algorithm 2. Induced Bisecting K-Means

1: Set the Intra Cluster Similarity (ICS) threshold parameter τ
2: Build a distance matrix A whose elements represents distance between couple of clips[2]
3: Select, as centroids, the two clips i and j s.t. $a_{ij} = \max_{l,m} A_{lm}$
4: Find 2 sub-clusters S1 and S2 using the basic k-Means algorithm
5: Check the ICS of S_1 and S_2 as
6: If the ICS value of a cluster is smaller than τ, then reapply the divisive process to this set, starting form step 2
7: If the ICS value of a cluster is over a given threshold, then stop. 6. The entire process will finish when there are no sub-clusters to divide.

The main differences of this algorithm with respect to the Standard Bisecting k- Means consist in: (1) how the initial centroids are chosen: as centroids of the two child clusters we select the clips of the parent cluster having the greatest distance between them; (2) the cluster splitting rule: a cluster is split in two subclusters if the Intra Cluster Similarity is smaller than the threshold value τ. Therefore, no input parameters K and ITER must be specified by the user. Our algorithm outputs a binary tree of clips, where each node represents a collection of similar clips. This dichotomic structure is then processed according to [20], in order to obtain a flat representation of clusters.

In order to perform an initial evaluation of the proposed multimedia summarization environmen, we considered 25 real proceedings characterized by a set of 3825 clips. A first analysis of the summary generated by our approach highlights two main peculiaries: high level of compression of the comprised multimedia documents, where two or three key frames per proceedings have been extracted, and high level of precision, i.e. the generated multimedia summaries contain the most important parts of the considered proceedings.

5 Conclusion and Future Work

In this paper a multimedia summarization environment has been presented in order to allow judicial actors to browse and navigate multimedia documents related to penal hearings/proceedings. The main component of this environment is represented by the summarization module, which creates a storyboard for the end user by exploiting several semantic information embedded into a courtroom recording. In particular, automatic speech transcriptions joint with automatic audio and video annotations have been used for deriving a compressed and meaningful representation of what happens into a law courtroom. Our work is now focused on creating a testing environment for a quality assessment of the produced storyboard.

[2] The distance metric is a linear combination of a cosine-based distance, for the numerical vectors concerned with transcriptions, and the jaccard distance, for the binary vectors concerned with audio/video annotations.

Acknowledgment

This work has been partially supported by the European Community FP-7 under the JUMAS Project (ref.: 214306).

References

1. Kim, J., Chang, H., Kang, K., Kim, M., Kim, H.: Summarization of news video and its description for content-based access. International Journal of Imaging Systems and Technology, 267–274 (2004)
2. Liang, C., Kuo, J., Chu, W., Wu, J.: Semantic units detection and summarization of baseball videos. In: Proc. of the 47th Midwest Symposium on Circuits and Systems, pp. 297–300 (2004)
3. Tjondronegoro, D.W., Chen, Y., Pham, B.: Classification of selfconsumable highlights for soccer video summaries. In: Proc. of the IEEE International Conference on Multimedia and Expo., pp. 579–582 (2003)
4. de Silva, G., Yamasaki, T., Aizawa, K.: Evaluation of video summarization for a large number of cameras in ubiquitous home. In: Proc. of the 13th Annual ACM International Conference on Multimedia, pp. 820–828 (2005)
5. Jaimes, A., Echigo, T., Teraguchi, M., Satoh, F.: Learning personalized video highlights from detailed MPEG-7 metadata. In: Proc. of the IEEE International Conference on Image Processing, pp. 133–136 (2002)
6. Takahashi, Y., Nitta, N., Babaguchi, N.: Video Summarization for Large Sports Video Archives. In: Proc. of the IEEE International Conference on Multimedia and Expo., pp. 1170–1173 (2005)
7. Agnihotri, L., Dimitrova, N., Kender, J.R.: Design and evaluation of a music video summarization system. In: Proc. of the IEEE International Conference on Multimedia and Expo., pp. 1943–1946 (2004)
8. Yang, H., Chaisorn, L., Zhao, Y., Neo, S., Chua, T.: VideoQA: question answering on news video. In: Proc. of the 11th Annual ACM International Conference on Multimedia, pp. 632–641 (2003)
9. Moriyama, T., Sakauchi, M.: Video summarization based on the psychological unfolding of drama. Systems and Computers in Japan, 1122–1131 (2002)
10. Rui, Y., Zhou, S.X., Huang, T.S.: Efficient access to video content in a unified framework. In: Proc. of the IEEE International Conference on Multimedia Computing and Systems, pp. 735–740 (1999)
11. Tseng, B.L., Smith, C.-Y.L.J.R.: Using MPEG-7 and MPEG-21 for personalizing video. IEEE Transactions on Multimedia, 42–52 (2004)
12. Steinbach, M., Karypis, G., Kumar, V.: A comparison of Document Clustering Techniques. In: KDD Workshop on Text Mining (2000)
13. Archetti, F., Fersini, E., Campanelli, P., Messina, E.: A Hierarchical Document Clustering Environment Based on the Induced Bisecting k-Means. In: Larsen, H.L., Pasi, G., Ortiz-Arroyo, D., Andreasen, T., Christiansen, H. (eds.) FQAS 2006. LNCS (LNAI), vol. 4027, pp. 257–269. Springer, Heidelberg (2006)
14. Lf, J., Gollan, C., Ney, H.: Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System. In: Interspeech, Brighton, U.K., September 2009, pp. 88–91 (2009)

15. Falavigna, D., Giuliani, D., Gretter, R., Lf, J., Gollan, C., Schlter, R., Ney, H.: Automatic Transcription of Courtroom Recordings in the JUMAS project. In: Proc. of the 2nd International Conference on ICT Solutions for Justice, Skopje, Macedonia (September 2009)
16. Avgerinakis, K., Briassouli, A., Kompatsiaris, I.: Video processing for judicial applications. In: Proc. of the 2nd International Conference on ICT Solutions for Justice, Skopje (2009)
17. Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
18. Fersini, E., Messina, E., Arosio, G., Archetti, F.: Audio-based Emotion Recognition in Judicial Domain: A Multilayer Support Vector Machines Approach. In: Perner, P. (ed.) Machine Learning and Data Mining in Pattern Recognition. LNCS, vol. 5632, pp. 594–602. Springer, Heidelberg (2009)
19. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988)
20. Kashyap, V., Ramakrishnan, C., Thomas, C., Bassu, D., Rindflesch, T.C., Sheth, A.: TaxaMiner: An experiment framework for automated taxonomy bootstrapping. International Journal of Web and Grid Services 1(2), 240–266 (2005)