A COMBINED THEMATIC AND ACOUSTIC APPROACH FOR A MUSIC RECOMMENDATION SERVICE IN TV COMMERCIALS

Mohamed Morchid, Richard Dufour, Georges Linarès
LIA - University of Avignon (France)
{mohamed.morchid, richard.dufour, georges.linares}@univ-avignon.fr

ABSTRACT

Most of modern advertisements contain a song to illustrate the commercial message. The success of a product, and its economic impact, can be directly linked to this choice. Finding the most appropriate song is usually made manually. Nonetheless, a single person is not able to listen and choose the best music among millions. The need for an automatic system for this particular task becomes increasingly critical. This paper describes the LIA music recommendation system for advertisements using both textual and acoustic features. This system aims at providing a song to a given commercial video and was evaluated in the context of the MediaEval 2013 Soundtrack task [14]. The goal of this task is to predict the most suitable soundtrack from a list of candidate songs, given a TV commercial. The organizers provide a development dataset including multimedia features. The initial assumption of the proposed system is that commercials which sell the same type of product, should also share the same music rhythm. A two-fold system is proposed: find commercials with close subjects in order to determine the mean rhythm of this subset, and then extract, from the candidate songs, the music which better corresponds to this mean rhythm.

1. INTRODUCTION

The success of a product or a service essentially depends of the way to present it. Thus, companies pay much attention to choose the most appropriate advertisement that will make a difference in the customer choice. The advertisers have different media possibilities, such as journal paper, radio, TV or Internet. In this context, they can exploit the audio media (TV, radio...) to attract listeners using a song related to the commercial. The choice of an appropriate song is crucial and can have a significant economic impact [5,18]. Usually, this choice is made by a human expert. Nonetheless, while millions of musics exist, a human agent could only choose a song among a limited subset. This choice could then be inappropriate, or simply not the best one, since the agent could not search into a large number of musics. For these reasons, the need for an automatic song recommandation system, to illustrate advertisements, becomes a critical subject for companies.

In this paper, an automatic system for songs recommandation is proposed. The proposed approach combines both textual (web pages) and audio (acoustic) features to select, among a large number of songs, the most appropriate and relevant music knowing the commercial content. The first step of the proposed system is to represent commercials into a thematic space built from a Latent Dirichlet Allocation (LDA) [4]. This pre-processing subtask uses the related textual content of the commercial. Then, acoustic features of each song are extracted to find a set of the most relevant songs for a given commercial.

An appropriate benchmark is needed to evaluate the effectiveness of the proposed recommandation system. For these reasons, the proposed system is evaluated in the context of the challenging MediaEval 2013 Soundtrack task for commercials [10]. Indeed, the MusiClef task seeks to make this process automated by taking into account both context- and content-based information about the video, the brand, and the music. The main difficulty of this task is to find the set of relevant features that best describes the most appropriate song for a video.

Next section describes related work in topic space modeling for information retrieval and music tasks. Section 3 presents the proposed music recommandation system using both textual content and acoustic features related to musics from commercials. Section 4 explains in details the unsupervised Latent Dirichlet Allocation (LDA) technique, while Section 4.2 describes how the acoustic features are used to evaluate the proximity of a music to a commercial. Finally, experiments are presented in Section 5, while Section 6 gives conclusions and perspectives.

2. RELATED WORKS

Latent Dirichlet Allocation (LDA) [4] is widely used in several tasks of information retrieval such as classification or keywords extraction. However, this unsupervised method is not much considered in the music processing tasks. Next sections describe related works using LDA techniques with text corpora (Section 2.1) and in the context of music tasks (Section 2.3).

2.1 Topic modeling

Several methods were proposed by Information Retrieval (IR) researchers to build topic spaces such as Latent Semantic Analysis or Indexing (LSA/LSI) [2, 6], that use a singular value decomposition (SVD) to reduce the space dimension.

This method was improved by [11] which proposed a probabilistic LSA/LSI (pLSA/pLSI). The pLSI approach models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of topics. This method demonstrated its performance on various tasks, such as sentence [3] or keyword [24] extraction. In spite of the effectiveness of the pLSI approach, this method has two main drawbacks. The distribution of topics in pLSI is indexed by training documents. Thus, the number of these parameters grows with the training document set size, and then, the model is prone to overfitting which is a main issue in an IR task such as document clustering. However, to address this shortcoming, a tempering heuristic is used to smooth the parameter of pLSI model for acceptable predictive performance. Nonetheless, authors showed in [20] that overfitting can occur even if tempering process is used.

As a result, IR researchers proposed the Latent Dirichlet allocation (LDA) [4] method to overcome these two drawbacks. Thus, the number of parameters of LDA does not grow with the size of the training corpus and LDA is not candidate for overfitting. LDA is a generative model which considers a document, seen as a *bag-of-words* [21], as a mixture of latent topics. In opposition to a multinomial mixture model, LDA considers that a theme is associated to each occurrence of a word composing the document, rather than associate a topic with the complete document. Thereby, a document can change of topics from a word to another. However, the word occurrences are connected by a latent variable which controls the global respect of the distribution of the topics in the document. These latent topics are characterized by a distribution of word probabilities which are associated with them. pLSI and LDA models have been shown to generally outperform LSI on IR tasks [12]. Moreover, LDA provides a direct estimate of the relevance of a topic knowing a word set or a document such as a web pages in the proposed system.

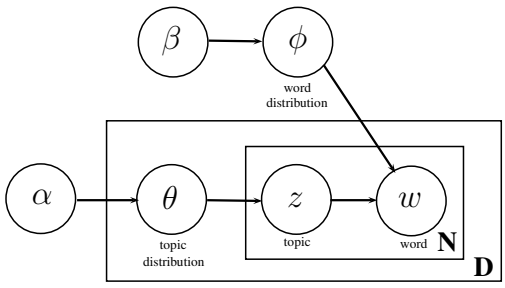

Figure 1. LDA Formalism.

Figure 1 presents the LDA formalism. For every document d of a corpus **D**, a first parameter θ is drawn according to a Dirichlet law of parameter α. A second parameter ϕ is drawn according to the same Dirichlet law of parameter β. Then, to generate every word w of the document d, a latent topic z is drawn from a multinomial distribution on θ. Knowing this topic z, the distribution of the words is a multinomial of parameters ϕ. The parameter θ is drawn for all the documents from the same *prior* parameter α. This allows to obtain a parameter binding the documents all together [4].

2.2 Gibbs sampling

Several techniques have been proposed to estimate LDA parameters, such as Variational Methods [4], Expectation-Propagation [17] or Gibbs Sampling [8]. Gibbs Sampling is a special case of Markov-chain Monte Carlo (MCMC) [7] and gives a simple algorithm for approximate inference in high-dimensional models such as LDA [9]. This overcomes the difficulty to directly and exactly estimate parameters that maximize the likelihood of the whole data collection defined as: $P(W|\overrightarrow{\alpha}, \overrightarrow{\beta}) = \prod_{m=1}^{M} P(\overrightarrow{w}_m|\overrightarrow{\alpha}, \overrightarrow{\beta})$ for the whole data collection $W = \{\overrightarrow{w}_m\}_{m=1}^{M}$ knowing the Dirichlet parameters $\overrightarrow{\alpha}$ and $\overrightarrow{\beta}$.

The first use of Gibbs Sampling for estimating LDA is reported in [8] and a more comprehensive description of this method can be found in [9]. One can refer to these papers for a better understanding of this sampling technique.

2.3 Topic modeling and Music

Topic modeling was already used in music processing, such as [13], where the authors presented a system which learns musical key as a key-profile. Thus, the proposed approach considered a song as a random mixture of key-profiles. In [25], authors described a classification method to assign a label to an unseen music. The authors use LDA to build a topic space from music-tags to get the probability of every music-tag belonging to each music genre. Then, each music is labeled to a genre knowing its tags. The purpose of the proposed approach is to find a set of relevant musics for a TV commercial.

3. PROPOSED APPROACH

The goal of the proposed automatic system is to recommend a set of musics given a TV commercial. The system uses external knowledge to find these songs. These external resources are composed with a set of TV commercials associated, for each one, with a song and a set of web pages (see [14] for more details about the MediaEval 2013 Soundtrack task). The idea behind the proposed approach is to assume that two commercials sharing same subjects or interests, also share the same kind of songs. The main issue in this approach is to find commercials, from the external dataset, that have sets of subjects close to those in commercials from the test set. As described in Section 2.1, a document can be represented as a set of latent topics. Thus, two documents sharing the same topics could be seen as *thematically* close.

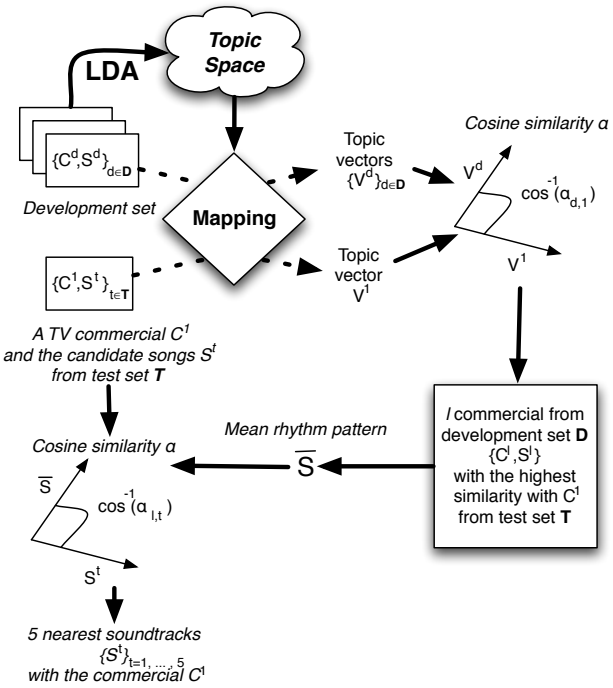

Figure 2. Global architecture of the proposed system.

Basically, the first process of the proposed three step system is to map each TV commercial from the test and development sets, into a topic space learnt with a LDA algorithm. A TV commercial from the test set is then linked to TV commercials from development set sharing a set of close topics. Moreover, each commercial of the development set is related to a music. Thus, as a result, a commercial from the test set is related to a subset of songs from the development set, considered as thematically close to the commercial textual content.

The second step has the responsibility to estimate a list of candidate songs (see Figure 2) using song audio features from the subset of songs thematically close associated during the first step. This subset of songs is used to evaluate a rhythm pattern of the *ideal* song for this commercial.

The last step retrieves, from all candidate songs from the test set, the closest song to the rhythm pattern estimated during the previous step.

In details, the development set **D** is composed of TV commercials C^d, with for each, a soundtrack S^d and a vector representation V^d related to the d^{th} TV commercial. In the same manner, the test set **T** is composed of TV commercials C^t, with, for the t^{th} one, a vector representation V^t and a soundtrack S^t to predict. Then a similarity score $\{\alpha_{d,t}\}_{d=1,\ldots,\mathbf{D}}^{t=1,\ldots,\mathbf{T}}$ is computed for each commercial C_i^d of the development set given one from the test set C^t:

$$\begin{aligned}\mathbf{D} &= \{C^d, V^D, S^d\}_{d=1,\ldots,\mathbf{D}} \\ \mathbf{T} &= \{C^t, V^T, S_k^t\}_{t=1,\ldots,\mathbf{T}}^{k=1,\ldots,5000}\end{aligned} \quad (1)$$

In the next sections, the topic space representation and the mapping of a commercial in this topic representation to evaluate both V^d and V^t are described. Then, the computed similarity score is detailed. Finally, the soundtrack prediction process from a TV commercial is explained.

4. TOPIC REPRESENTATION OF A TV COMMERCIAL

Let's consider a corpus **D** from the development set of TV commercials with a word vocabulary $\mathbf{V} = \{w_1, \ldots, w_N\}$ of size N. A topic representation from corpus **D** is then performed using a Latent Dirichlet Allocation (LDA) [4] approach. At the final LDA analysis, a topic space m of n topics is obtained with, for each theme z, the probability of each word w of **V** knowing z, and for the entire model m, the probability of each theme z knowing the model m. Each TV commercial from both development and test sets is mapped into the topic space (see Figure 3) to obtain a vector representation (V^d and V^t) of web pages related to a commercial into the thematic space computed as follow:

$$V^d[i](C_j^d) = P(z_i|C_j^d) \quad (2)$$

where $P(z_i|C_j^d)$ is the probability of a topic z_i to be generated by the web pages from the commercial C_j^d, estimated using Gibbs sampling as described in Section 2.2. In the same way, V^t is estimated with the same topic space, and with the use of web pages of commercials of test set C_j^t (see Figure 3).

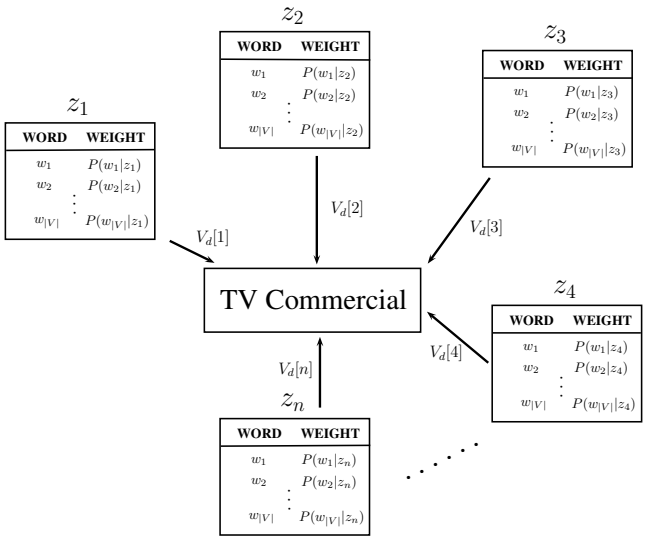

Figure 3. Mapping of a TV commercial in the topic space.

4.1 Similarity measure

Each commercial from both development and test set, is mapped into the topic space to produce a vector representation for each one, respectively V^d and V^t as outcomes. Then, given a TV commercial C^1 from the test set **T**, a subset of other TV commercials from the development set **D** is selected knowing their thematic proximity with C^1.

To estimate the similarity between C^1 and commercials from development set, the cosine metric α is used. This similarity metric is expressed thereafter:

$$cosine(V^d, V^t) = \alpha_{d,t} = \frac{\sum_{i=1}^{n} V^d[i] \times V^t[i]}{\sqrt{\sum_{i=1}^{n} V^d[i]^2}\sqrt{\sum_{i=1}^{n} V^t[i]^2}} \quad (3)$$

This metric allows to extract a subset of commercials from **D** thematically close to C^1.

4.2 Rhythm pattern

The cosine measure, presented in previous section, is also used to evaluate the similarity between a mean rhythm pattern vector $\overline{S^d}$ of a song, and all the candidate songs S_k^t of the test set.

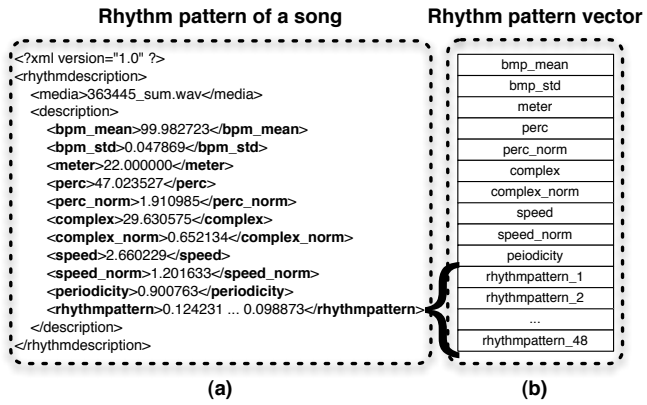

Figure 4. Rhythm pattern of a song from the development set in xml (a) and vector (b) representations.

In details, each commercial from **D** is related with a soundtrack that is represented with a rhythm pattern vector. The organizers provide for each song contained into the MusicClef 2013 dataset:

- *video features* (MPEG-7 Motion Activity and Scalable Color Descriptor [15]),
- web pages about the respective brands and music artists,
- *music features*:
  - MFCC or BLF [22],
  - PS209 [19],
  - beat, key, harmonic pattern extracted with the Ircam software [1].

In our experiments, 10 rhythm features of songs are used (*speed*, *percussion,* ..., *periodicity*) as shown in Figure 4. These features of beat, key or harmonic pattern are extracted using the Ircam software available at [1]. More information about features extraction from songs are detailed in [14].

As an outcome, each commercial is represented by a rhythm pattern vector of size 58 (10 from song features and 48 from rhythm pattern). From the subset of soundtracks of the l nearest commercials from **D**, a mean rhythm vector $\overline{S}$ is performed as:

$$\overline{S} = \frac{1}{l}\sum_{d \in l} S^d \,.$$

Finally, the cosine measure between this mean rhythm $\overline{S}$ of the l nearest commercials from **D**, and each commercial ($cosine(\overline{S}, S^t)_{t \in T}$), is used to find, from the soundtrack S^t of the test set **T**, the 5 songs from all the candidates having the closest rhythm pattern.

5. EXPERIMENTS AND RESULTS

Previous sections described the proposed automatic music recommandation system for TV commercials. This system is decomposed into three sub-processes. The first one maps the commercials into a topic space to evaluate the proximity of a commercial from the test set and all commercials from the development set. Then, the mean rhythm pattern of the thematically close commercials is computed. Finally, this rhythm pattern is computed with all ones from the test set of candidate songs to find a set of relevant musics.

5.1 Experimental protocol

The first step of the proposed approach, detailed in previous section, maps TV commercial textual content into a topic space of size n ($n = 500$). This one is learnt from a LDA in a large corpus of documents. Section 4 describes the corpus **D** of web pages. This corpus contains $10,724$ Web pages related to brands of the commercials contained in **D**. This corpus is composed of $44,229,747$ words for a vocabulary of $4,476,153$ unique words. More details about this text corpus, and the way to collect it, is explained into [14].

The first step of the proposed approach is to map each commercial textual content into a topic space learnt from a latent Dirichlet allocation (LDA). During the experiments, the MALLET tool is used [16] to perform a topic model. The proposed system is evaluated in the MediaEval 2013 MusiClef benchmark [14]. The aim of this task is to predict, for each video of the test set, the most suitable soundtrack from 5,000 candidate songs. The dataset is split into 3 sets. The development set contains multimodal information on 392 commercials (various metadata including Youtube uploader comments, audio features, video features, web pages and text features). The test set is a set of 55 videos to which a song should be associated using the recommandation set of 5,000 soundtracks (30 seconds long excerpts).

5.2 Experimental metrics

For each video in the test set, a ranked list of 5 candidate songs should be proposed. The song prediction evaluation is manually performed using the Amazon Mechanical Turk platform. This novel task is non-trivial in terms of "ground truth", that is why human ratings for evaluation are used. Three scores have been computed from our system output. Let V be the full collection of test set videos, and let $\overline{s_r(v)}$ be the average suitability score for the audio file suggested at rank r for the video v. Then, the evaluation measures are computed as follows:

- Average suitability score of the first-ranked song: $\frac{1}{V}\sum_{i=1}^{|V|}\overline{s_1(v_i)}$
- Average suitability score for the full top-5: $\frac{1}{V}\sum_{i=1}^{|V|}\frac{1}{5}\overline{s_r(v_i)}$
- Weighted average suitability score of the full top-5. Here, we apply a weighted harmonic mean score instead of an arithmetic mean: $\frac{1}{V}\sum_{i=1}^{|V|}\frac{\sum_{r=1}^{5}\overline{s_r(v_i)}}{\sum_{r=1}^{5}\frac{\overline{s_r(v_i)}}{r}}$

The previously presented measures are used to study both rating and ranking aspects of the results.

5.3 Results

The measures defined in the previous section are used to evaluate the effectiveness of songs selected to be associated to TV commercials from the test set. The proposed topic space-based approach is evaluated in the same way, and obtained the results detailed thereafter:

- First rank average score: **2.16**
- Top 5 average score (arithmetic mean): **2.24**
- Top 5 average score (harmonic mean, taking rank into account): **2.22**

Considering that human judges rate the predicted songs from 1 (*very poor*) to 4 (*very well*), we can consider that our system is slightly better than the mean evaluation score (2) no matter the metric considered. While the system proposed in [23] is clearly different from ours, results are very similar. This shows the difficulty to build an automatic song recommendation system for TV commercials, the evaluation being also a critical point to discuss.

6. CONCLUSIONS AND PERSPECTIVES

In this paper, an automatic system to assign a soundtrack to a TV commercial has been proposed. This system combines two media: textual commercial content and audio rhythm pattern. The proposed approach obtains good results in spite of the fact that the system is automatic and unsupervised. Indeed, both subtasks are unsupervised (LDA learning and commercials mapping into the topic space) and songs extraction (rhythm pattern estimation of the *ideal* songs for a commercial from the test set). Moreover, this promising approach, combining thematic representation of the textual content of a set of web pages describing a TV commercial and acoustic features, shows the relevance of topic-based representation in automatic recommandation using external resources (development set).

The choice of a relevant song to describe the idea behind a commercial, is a challenging task when the framework does not take into account relevant features related to:

- mood, such as harmonic content, harmonic progressions and timbre,
- music rhythm, such as musical style, texture, spectral centroid, or tempo.

The proposed automatic music recommendation system is limited by this small number (58) of features which not describe all music aspects. For these reasons, in future works, we plan to use others features, such as the song lyrics or the audio transcription of the TV commercials, and evaluate the effectiveness of the proposed hybrid framework into other information retrieval tasks such as classification of music genre or music clustering.

7. REFERENCES

[1] Ircam. analyse-synthse: Software. In *http://anasynth.ircam.fr/home/software.*, Accessed: Sept. 2013.

[2] J.R. Bellegarda. A latent semantic analysis framework for large-span language modeling. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[3] J.R. Bellegarda. Exploiting latent semantic information in statistical language modeling. *Proceedings of the IEEE*, 88(8):1279–1296, 2000.

[4] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Claudia Bullerjahn. The effectiveness of music in television commercials. *Food Preferences and Taste: Continuity and Change*, 2:207, 1997.

[6] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[7] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):721–741, 1984.

[8] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[9] Gregor Heinrich. Parameter estimation for text analysis. *Web: http://www. arbylon. net/publications/text-est. pdf*, 2005.

[10] Nina Hoeberichts. Music and advertising: The effect of music in television commercials on consumer attitudes. *Bachelor Thesis*, 2012.

[11] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI ' 99*, page 21. Citeseer, 1999.

[12] Thomas Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[13] Diane Hu and Lawrence K Saul. A probabilistic topic model for unsupervised learning of musical key-profiles. In *ISMIR*, pages 441–446, 2009.

[14] Cynthia C. S. Liem, Nicola Orio, Geoffroy Peeters, and Markus Scheld. MusiClef 2013: Soundtrack Selection for Commercials. In *MediaEval*, 2013.

[15] Bangalore S Manjunath, Philippe Salembier, and Thomas Sikora. *Introduction to MPEG-7: multimedia content description interface*, volume 1. John Wiley & Sons, 2002.

[16] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[17] Thomas Minka and John Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 352–359. Morgan Kaufmann Publishers Inc., 2002.

[18] C Whan Park and S Mark Young. Consumer response to television commercials: The impact of involvement and background music on brand attitude formation. *Journal of Marketing Research*, pages 11–24, 1986.

[19] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer. On rhythm and general music similarity. In *ISMIR*, pages 525–530, 2009.

[20] Alexandrin Popescul, David M Pennock, and Steve Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 437–444. Morgan Kaufmann Publishers Inc., 2001.

[21] G. Salton. Automatic text processing: the transformation. *Analysis and Retrieval of Information by Computer*, 1989.

[22] Klaus Seyerlehner, Gerhard Widmer, and Tim Pohle. Fusing block-level features for music similarity estimation. In *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, pages 225–232, 2010.

[23] Han Su, Fang-Fei Kuo, Chu-Hsiang Chiu, Yen-Ju Chou, and Man-Kwan Shan. Mediaeval 2013: Soundtrack selection for commercials based on content correlation modeling. In *MediaEval 2013*, volume 1043 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[24] Y. Suzuki, F. Fukumoto, and Y. Sekiguchi. Keyword extraction using term-domain interdependence for dictation of radio news. In *17th international conference on Computational linguistics*, volume 2, pages 1272–1276. ACL, 1998.

[25] Chao Zhen and Jieping Xu. Multi-modal music genre classification approach. In *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, volume 8, pages 398–402. IEEE, 2010.