Accepted version of paper published in the 9th International Conference on Image Analysis and Processing (ICIAP), 2017.
DOI: https://doi.org/10.1007/978-3-319-68560-1

A Rank Aggregation Framework for Video Interestingness Prediction*

Jurandy Almeida[1], Lucas P. Valem[2], and Daniel C. G. Pedronette[2]

[1] Institute of Science and Technology, Federal University of São Paulo – UNIFESP
12247-014, São José dos Campos – Brazil
jurandy.almeida@unifesp.br
[2] Dept. Stat., Appl. Math. & Comput., State University of São Paulo – UNESP
13506-900, Rio Claro – Brazil
{lucasvalem,daniel}@rc.unesp.br

Abstract. Often, different segments of a video may be more or less attractive for people depending on their experience in watching it. Due to this subjectiveness, the challenging task of automatically predicting whether a video segment is interesting or not has attracted a lot of attention. Current solutions are usually based on learning models trained with features from different modalities. In this paper, we propose a late fusion with rank aggregation methods for combining ranking models learned with features of different modalities and by different learning-to-rank algorithms. The experimental evaluation was conducted on a benchmarking dataset provided for the Predicting Media Interestingness Task at the MediaEval 2016. Two different modalities and four learning-to-rank algorithms are considered. The results are promising and show that the rank aggregation methods can be used to improve the overall performance, reaching gains of more than 10% over state-of-the-art solutions.

Keywords: multimedia information retrieval, predicting media interestingness, learning-to-rank methods, multimodal late fusion, rank aggregation

1 Introduction

The production of multimedia data have been grown continuously and consistently. Supported by mobile devices, social networks and cloud environments, multimedia data can be generated, shared and stored everywhere. In this scenario, there is a growing demand for efficient systems able to manage large volumes of multimedia data and reduce the work and information overload when seeking a given content of interest [18].

However, several research challenges are involved, from content representation to its indexing and ranking according to user interests, specially considering different modalities. In many multimedia applications, the fusion of different modalities is essential for improving the overall performance [19, 23]. The main motivation of fusion approaches consists in achieving a more precise representation of the data by combining features from distinct modalities, such as audio and

* Thanks to Brazilian agencies FAPESP (grants 2013/08645-0, 2016/06441-7, and 2017/02091-4), CNPq (grant 423228/2016-1), and CAPES for funding.

visual content [20]. Additionally, different learning models capable of encoding user preferences can be also considered and fused as complementary information.

In this paper, a multimodal fusion framework based on rank aggregation is proposed for video interestingness prediction. Firstly, different audio and visual features are extracted for constructing a content-based representation. Subsequently, user preferences are encoded through learning-to-rank algorithms, used to construct rankers capable of predicting the interestingness degree of a video. Finally, rank aggregation methods are used for combining the multimodal information provided by different pairs of feature-rankers in order to improve the effectiveness of predictions. Experimental results demonstrate the potential of rank aggregation methods for combining multimodal information on interestingness prediction tasks, which can improve the state-of-the-art results [1] in more than 10%. In addition, the relevance of feature selection strategy is also discussed, providing useful guidance for future work.

This paper is organized as follows. Section 2 discusses related work. Section 3 presents the features, while Section 4 presents the learning-to-rank algorithms. Section 5 discusses the rank aggregation methods. Section 6 reports the results of our experiments. Finally, Section 7 states conclusions and presents future work.

2 Related Work

This section presents an overview of related work dedicated to video interestingness prediction. In this work, we are interested in multimodal approaches based on data fusion.

The pioneering work of Jiang et al. [16] introduced a new dataset for predicting the interestingness of videos, where a large number of features were evaluated and used to train prediction models with Ranking SVM [17]. According to their findings, audio and visual features are effective for approaching this task, and their fusion can improve the overall performance.

A lot of research on video interestingness prediction has been done for the MediaEval 2016 Predicting Media Interestingness Task [12]. This task aims to automatically select the most interesting video shots according to a common viewer by using features derived from audio-visual content or associated textual information. Ten groups submitted their results for the video subtask and six of them adopted a multimodal approach. The final ranking of these six groups based on the official results was: RECOD [1], UNIGECISA [26], RUC [8], NII-UIT [22], Technicolor [27], and BigVid [29].

Almeida [1] (RECOD team) extracted motion features from the video shots and used them to train four different ranking models, which were combined by a majority voting strategy. Here, we extend the work of Almeida by exploring data fusion (audio and visual data) to enhance video interestingness prediction.

Rayatdoost and Soleymani [26] (UNIGECISA team) used both audio and keyframe-based features provided for the task. Also, they extracted visual sentiment and emotional acoustic features. To obtain a single representation for each shot, they computed the mean and the standard deviation for all the keyframes. Then, principal component analysis (PCA) were applied to reduce the dimensionality of such features. Finally, three different regression models were trained based on the reduced features.

Chen et al. [8] (RUC team) used both audio and keyframe-based features provided for the task. In addition, they extracted statistical acoustic and deep learning features. A single representation for each shot was computed by applying mean pooling over all the keyframe-based features. Different features were combined by early fusion and used to train two different classification models.

Lan et al. [22] (NII-UIT team) used both audio and keyframe-based features provided for the task and also extracted deep learning features. A max pooling strategy was used to aggregate all the keyframe-based features into a single representation for each shot, which was used to train a SVM (Support Vector Machine) classifier. Classification models learned with different features are combined by late fusion using an average weighting scheme.

Shen et al. [27] (Technicolor team) used both audio and keyframe-based features provided for the task. They used such features to train two different deep neural network architectures.

Xu et al. [29] (BigVid team) used both audio and keyframe-based features provided for the task. Also, they extracted semantic features based on sentiment and style attributes. Average pooling over all the keyframe-based features was applied to compute a single representation for each shot. Such features were used to train three different learning models: a classification model using SVM, a ranking model using Ranking SVM, and a deep neural network. In addition, they also considered the combination between SVM and Ranking SVM using a score-level average late fusion.

In this work, we propose a late fusion with rank aggregation methods for combining ranking models learned with features of different modalities and by different learning-to-rank algorithms.

3 Feature Extraction

Two main approaches were used to encode video content. One of them encodes motion information by using *histogram of motion patterns* [2]. The other approach is based on audio information and considers the well-known *mel-frequency cepstral coefficients* [11].

3.1 Histogram of Motion Patterns

Instead of using any keyframe visual features, a simple and fast algorithm was adopted to encode visual properties, known as *histogram of motion patterns* (HMP) [2]. It considers the video movement by the transitions between frames. For each frame of an input video, motion features are extracted from the video stream. For that, 2×2 ordinal matrices are obtained by ranking the intensity values of the four luminance (Y) blocks of each macro block. This strategy is employed for computing both the spatial feature of the 4-blocks of a macro block and the temporal feature of corresponding blocks in three frames (previous, current, and next). Each possible combination of the ordinal measures is treated as an individual pattern of 16-bits (i.e., 2-bits for each element of the ordinal matrices). Finally, the spatio-temporal pattern of all the macro blocks of the video sequence are accumulated to form a normalized histogram.

3.2 Mel-Frequency Cepstral Coefficients

Besides encoding visual properties using HMP, we also used a representation very popular to encode audio information, called *mel-frequency cepstral coefficients* (MFCC) [11]. They are capable of representing the short-time power spectrum of a sound in an accurate and compact form. Initially, the audio signal is filtered with a Finite Impulse Response (FIR) filter to pre-amplify high frequencies. Then, the resulting signal is converted to frames of small duration (typically 20-40ms). Next, such frames are weighted by a Hamming window aiming at removing any negative effects on its edges. After that, the power spectrum of each frame is computed by applying the Discrete Fourier Transform (DFT) and taking only the magnitude of the spectral coefficients. Thereafter, a filter bank of overlapping triangular filters, also known as Mel-scale filter bank, is used to smooth the spectrum and emphasize perceptually meaningful frequencies. Once the filterbank energies are computed, the logarithm of them is taken aiming at reducing large variations in energy, whose loudness is not perceived by humans. Finally, the Discrete Cosine Transform (DCT) is applied to the log Mel filterbank energies and then only the lower-order coefficients are used to form the feature vector.

4 Ranking Models

The interestingness of videos is a subjective concept that depends on judgments of different viewers on whether a video is interesting or not based on their experience in watching it. Due to this subjectiveness, the automatic prediction of the interestingness degree of a video is a challenging task.

To approach this task, we adopted the strategy proposed by Jiang et al. [16], where a machine learning model is trained aiming at comparing the interestingness between video pairs. In this way, given two videos to the system, it indicates the more interesting one. The basic idea is to use machine learning algorithms to learn a ranking function based on features extracted from training data, and then apply it to features extracted from testing data.

We have used four different learning-to-rank algorithms. The first three are based on pairwise comparisons: *Ranking SVM* [17], *RankNet* [6], and *RankBoost* [14]. The latter approach considers lists of objects by using *ListNet* [7].

Ranking SVM [17] is a pairwise ranking method that uses the Support Vector Machine (SVM) classifier to learn a ranking function. For that, each query and its possible results are mapped to a feature space. Next, a given rank is associated to each point in this space. Finally, a SVM classifier is used to find an optimal separating hyperplane between those points based on their ranks.

RankNet [6] is a pairwise ranking method that relies on a probabilistic model. For that, pairwise rankings are transformed into probability distributions, enabling the use of probability distribution metrics as cost functions. Thus, optimization algorithms can be used to minimize a cost function to perform pairwise rankings. The authors formulate this cost function using a neural network in which the learning rate is controlled with gradient descent steps.

RankBoost [14] is a pairwise ranking method that relies on boosting algorithms. Initially, each possible result for a given query is mapped to a feature

space, in which each dimension indicates the relative ranking of individual pairs of results, i.e., whether one result is ranked below or above the other. Thus, the ranking problem is formulated as a binary classification problem. Next, a set of weak rankers are trained iteratively. At each iteration, the resulting pairs are re-weighted so that the weight of pairs ranked wrongly is increased whereas the weight of pairs ranked correctly is decreased. Finally, all the weak rankers are combined as a final ranking function.

ListNet [7] is an extension of RankNet that, instead of using pairwise rankings, considers all possible results for a given query as a single instance, enabling to capture and exploit the intrinsic structure of the data. Roughly speaking, it is a listwise ranking method that relies on the probability distribution of permutations. Initially, a given scoring function is used to define the permutation probability distribution for the predicted rankings. Then, another permutation probability distribution is defined for the ground truth. Next, the K-L divergence is used to compute the cross entropy between these two distributions, which is defined as the listwise ranking loss between them. Finally, a linear neural network model is trained through the gradient descent algorithm, which is used to minimize the listwise ranking loss.

5 Rank Aggregation Framework

Ranking has been established as a relevant task in many diverse domains, including information retrieval, natural language processing, and collaborative filtering [9]. However, in many situations, distinct ranking models produce different results. Additionally, the information provided by different ranking results is often complementary, and therefore, can be used for improving the effectiveness of the systems. This is the objective of rank aggregation methods, which aim at combining different rankings in order to obtain a more accurate one.

Rank aggregation approaches are often unsupervised, requiring no training data and can be seen as a way for obtaining a consensus ranking when multiple scores or ranked lists are provided for a set of objects. Different strategies have been used, considering mainly the information of the score computed for an object and the position (or rank) assigned to an object in a ranked list.

Formally, a rank aggregation method can be defined as follows. Let $\mathcal{C}$=$\{vs_1, vs_2, \ldots, vs_n\}$ be a collection of video shots, where n denotes the number of shots for the video being analyzed. Let $\mathcal{D}$=$\{D_1, D_2, \ldots, D_d\}$ be a set of rankers. Let the function $\rho_j(i)$ denotes the interestingness degree assigned by the ranker $D_j \in \mathcal{D}$ to the video shot $vs_i \in \mathcal{C}$.

Based on the score ρ_j, a ranked list τ_j=$(vs_1, vs_2, \ldots, vs_n)$ can be computed. The ranked list τ_j can be defined as a permutation of the collection $\mathcal{C}$, which contains the most interesting video shots according to the ranker D_j. A permutation τ_j is a bijection from the set $\mathcal{C}$ onto the set $[n] = \{1, 2, \ldots, n\}$. For a permutation τ_j, we interpret $\tau_j(i)$ as the position (or rank) of the video shot vs_i in the ranked list τ_j. We can say that, if vs_i is ranked before vs_l in the ranked list τ_j, that is, $\tau_j(i) < \tau_j(l)$, then $\rho(j, i) \geq \rho(j, l)$.

Given the different scores ρ_j and their respective ranked lists τ_j computed by distinct rankers $D_j \in \mathcal{D}$, a rank aggregation method aims to compute a fused

score $F(i)$ to each video shot vs_i. In this work, we used three different methods based on score and rank information, described in the following sections.

5.1 Borda Method

The Borda [30] method combines the rank information of each video shot in different ranked lists computed by different rankers. The Borda count method uses rank information in voting procedures. Rank scores are linearly assigned to video shots in ranked lists according to their positions and are summed directly.

More specifically, the distance is scored by the number of video shots not ranked higher than it in the different ranked lists [21]. The new score $F_B(i)$ is computed as follows:

$$F_B(i) = \sum_{j=0}^{d} \tau_j(i). \tag{1}$$

5.2 Reciprocal Rank Fusion

The Reciprocal Rank Fusion [10] uses the rank information for computing a new score according to a naive scoring formula:

$$F_R(i) = \sum_{j=0}^{d} \frac{1}{k + \tau_j(i)}, \tag{2}$$

The intuition behind the formula is based on the conjecture that highly-ranked shots are significantly more relevant than lower-ranked shots [10]. The constant k mitigates the impact outlier rankers. For the experiments in this paper, $k = 16$ is used.

5.3 Multiplicative Rank Aggregation

A multiplicative approach [24] is used for the rank aggregation based on scores. The use of a multiplication approach is inspired by the Naïve Bayes classifiers. Given a set of scores computed by distinct rankers, such classifiers try to estimate the relevance probability assuming conditional independence among rankers. Considering the independence assumption, the scores of each ranker are multiplied. The fused score $F_M(i)$ for a given video shot vs_i is computed as:

$$F_M(i) = \prod_{j=1}^{d} (1 + \rho(j, i)). \tag{3}$$

6 Experiments and Results

Experiments were conducted on a benchmarking dataset provided by the MediaEval 2016 organizers for the Predicting Media Interestingness Task [12]. This dataset is composed of 78 Creative Commons licensed trailers of Hollywood-like movies. It is divided into a *development set* of 52 videos (67%) and a *test set*

of 26 videos (33%). These videos were segmented by hand, producing a total of 7,396 video shots. After video segmentation, the development set has 5,054 shots and the test set has 2,342 shots.

Each video shot was represented by the HMP and MFCC features, as discussed in Section 3. For encoding visual properties, we extracted the HMP features directly from the video data. On other hand, for representing audio information, we used the MFCC features provided for the task [15]. Unlike HMP, MFCC produces multiple local features for a same video. To obtain a single representation, we built a Bag-of-Features (BoF) [5] model upon local MFCC features. In the BoF framework, visual words [28] are obtained by quantizing local features according to a pre-learned dictionary. Thus, a video sequence is represented as a normalized frequency histogram of visual words associated with each local feature. In this work, we construct a codebook of 4000 visual words using a random selection. For the dictionary creation, we used only the MFCC features extrated from the development set.

Once the features were extracted, they were used as input to train machine-learned rankers, as presented in Section 4. The SVM^{rank} package[3] [17] was used for running Ranking SVM. The RankLib package[4] was used for running RankNet, RankBoost, and ListNet. Ranking SVM was configured with a linear kernel. RankNet, RankBoost, and ListNet were configured with their default parameter settings. All those approaches were calibrated through a 4-fold cross validation on the development set. Next, the trained rankers were used to predict the rankings of test video shots. The rankings associated with the video shots of a same movie trailer were normalized using a z-score normalization. After that, the normalized rankings of all the rankers are combined using our proposed framework, producing the final prediction scores. Finally, a thresholding method was applied to transform the prediction scores into binary decisions. It was found empirically that better results were obtained when a video shot is classified as interesting if its prediction score is greater than 0.7; otherwise, it is classified as non interesting.

The effectiveness of our strategy was assessed using Mean Average Precision (MAP), which is the official evaluation metric adopted in the task. Our results were compared with those reported by Almeida[5] [1], which ranked 1st out of 10 groups in the MediaEval 2016 Predicting Media Interestingness Task.

Table 1 presents the results obtained by the HMP and MFCC features in isolation. On the development set, by analyzing the confidence intervals, it can be noticed that the performanceof the different learning-to-rank algorithms is similar, with a small advantage to Ranking SVM. On the test set, however, Ranking SVM provided the best results for HMP whereas ListNet was the best for MFCC. These results indicate that the fusion of such learning-to-rank algorithms may be promising.

For combining the results provided by different features and machine-learned rankers, we adopted the strategy proposed by Almeida et al. [3]. Initially, we

[3] `https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

[4] `https://sourceforge.net/p/lemur/wiki/RankLib/`

[5] The results reported by Almeida [1] refer to those obtained using only the HMP feature and are presented in Table 1.

Table 1. Results obtained by HMP and MFCC on the development set using the machine-learned rankers in isolation.

Feature	Ranker	Development Set			Test Set
		Avg.	Conf. Interval (95%)		
			min.	*max.*	
HMP	Ranking SVM	15.19	13.99	16.38	18.15
	RankNet	13.82	12.09	15.55	16.17
	RankBoost	14.67	12.93	16.42	16.17
	ListNet	13.32	12.06	14.57	16.56
MFCC	Ranking SVM	14.19	12.27	16.12	15.87
	RankNet	13.33	11.49	15.17	17.10
	RankBoost	12.53	11.55	13.51	15.62
	ListNet	13.45	12.20	14.71	17.57

sorted the individual results obtained by each pair (feature & ranker) in a decreasing order of MAP. Then, each pair was selected according to its rank, i.e., the best was the first, the second best was the second, and so on. At each step, the next pair was combined with all the previous ones, as discussed in Section 5.

Figure 1 shows the MAP scores obtained by different rank aggregation methods on the development set. We show the behavior of such methods for combining the most effective pairs according to the average individual results achieved in the development set (see Table 1). The horizontal line denotes the MAP score for the best pair in isolation and forms a baseline for our proposed framework. The vertical line indicates the set of pairs which achieved the highest MAP score when combined with the rank aggregation methods. The error bars represent 95% confidence intervals computed from the 4 folds.

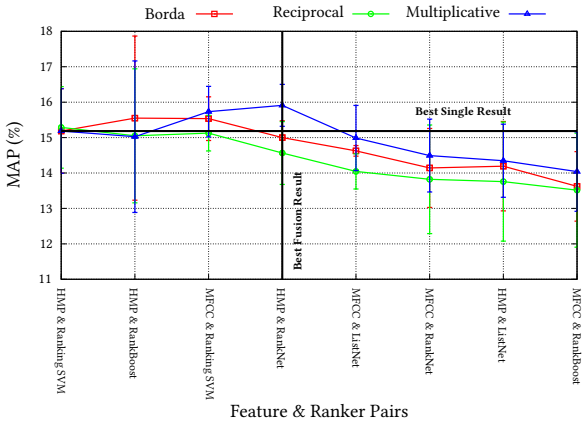
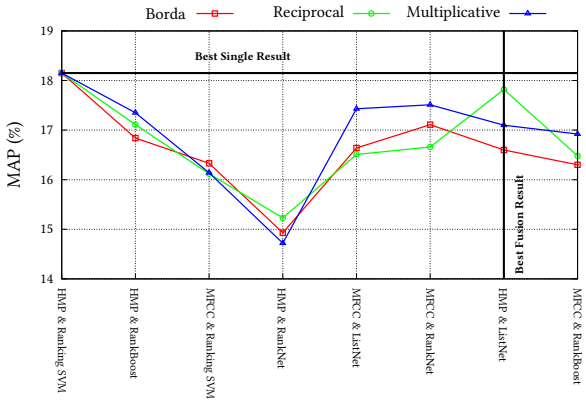


Fig. 1. MAP obtained by different rank aggregation methods on the development set.

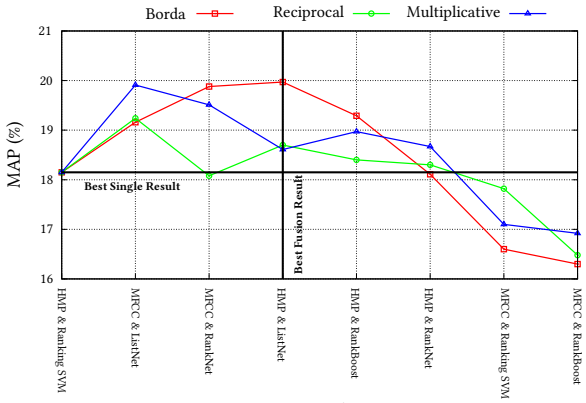
We can see that, as more pairs are considered for late fusion, more effective results are obtained, until reach a peak. This is an expected behavior, because different features and machine-learned rankers may complement each other, which aggregates more information. From a certain point, however, non-relevant results

from the less effective pairs exceed relevant results from the most effective ones and the gain decreases. By analyzing the confidence intervals, it is important to note that there is a high variance among the 4 folds. These results indicate that the ordering defined by such folds, i.e., from the most to the least effective pairs, is not consistent. This ordering is used for selecting the pairs to be combined by the rank aggregation methods.

Figure 2 shows the MAP scores obtained by different rank aggregation methods on the test set. In Figure 2(a), features and machine-learned rankers were selected for late fusion with rank aggregation methods in a decreasing order of their average individual results on the development set (see Table 1). Notice that the rank aggregation methods did not improve the best individual result (i.e., HMP & Ranking SVM). The main reason for such results is the selection strategy adopted for defining the pairs to be used for combination.



(a) Order defined by the Development Set



(b) Order defined by the Test Set

Fig. 2. MAP obtained by different rank aggregation methods on the test set.

In Figure 2(b), we replicate the previous experiment, however a different selection strategy was adopted. In this figure, we show the MAP scores as the

most effective pairs are used for combination. Unlike the previous experiment, instead of considering the decreasing order of average individual results from the development set, the ordering was defined based on the individual results achieved in the test set (see Table 1). As we can see, the best fusion result was obtained by the Borda method in combining the four most effective pairs (i.e., HMP & Ranking SVM, MFCC & ListNet, MFCC & RankNet, HMP & ListNet), which achieved a MAP score equals to **19.97%**, yielding gains of more than **10%** with respect to the best single result (i.e., HMP & Ranking SVM).

Such positive results indicate the potential of rank aggregation methods for combining multimodal information and improving the interestingness prediction. At the same time, the importance of the selection strategy is also evident. The better results presented by the set of pairs defined by the effectiveness order on the test set indicate that unsupervised selection procedures can be exploited.

7 Conclusions

This paper presented a novel approach for predicting the interestingness of videos. Our method is based on combining the features of audio and visual modalities with rank aggregation methods. The proposed strategy relies on a late fusion of ranking models learned with different learning-to-rank algorithms.

Our approach was validated in the dataset of the MediaEval 2016 Predicting Media Interestingness Task. Conducted experiments demonstrate that our multimodal strategy yields better video interestingness prediction results when compared with those based on a single modality (either audio or visual information). Also, we show that, by using a proper selection strategy, the rank aggregation methods can be used to improve the overall performance, achieving significant gains in comparison with state-of-the-art solutions.

Future work includes the evaluation of other features (e.g., keyframe-based methods [13, 25]), especially those encoding information from different modalities, as well as perform an extensive study on smarter selection strategies for combining learning-to-rank algorithms (e.g., genetic programming [4]).

References

1. Almeida, J.: UNIFESP at Mediaeval 2016: Predicting Media Interestingness task. In: Proc. of the MediaEval 2016 Workshop (2016)
2. Almeida, J., Leite, N.J., Torres, R.S.: Comparison of video sequences with histograms of motion patterns. In: ICIP. pp. 3673–3676 (2011)
3. Almeida, J., Pedronette, D.C.G., Alberton, B., Morellato, L.P.C., Torres, R.S.: Unsupervised distance learning for plant species identification. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 9(12), 5325–5338 (2016)
4. Andrade, F.S.P., Almeida, J., Pedrini, H., da S. Torres, R.: Fusion of local and global descriptors for content-based image and video retrieval. In: CIARP. pp. 845–853 (2012)
5. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: CVPR. pp. 2559–2566 (2010)
6. Burges, C.J.C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., Hullender, G.N.: Learning to rank using gradient descent. In: ICML. pp. 89–96 (2005)

7. Cao, Z., Qin, T., Liu, T.Y., Tsai, M.F., Li, H.: Learning to rank: from pairwise approach to listwise approach. In: ICML. pp. 129–136 (2007)
8. Chen, S., Dian, Y., Jin, Q.: RUC at Mediaeval 2016: Predicting Media Interestingness task. In: Proc. of the MediaEval 2016 Workshop (2016)
9. Chen, W., Liu, T.Y., Lan, Y., Ma, Z., Li, H.: Ranking measures and loss functions in learning to rank. In: NIPS. pp. 315–323 (2009)
10. Cormack, G.V., Clarke, C.L.A., Buettcher, S.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: SIGIR. pp. 758–759 (2009)
11. Davis, S., Mermelstein, P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust., Speech, Signal Process. 28(4), 357–366 (1980)
12. Demarty, C.H., Sjöberg, M., Ionescu, B., Do, T.T., Wang, H., Duong, N.Q.K., Lefebvre, F.: Mediaeval 2016 Predicting Media Interestingness task. In: Proc. of the MediaEval 2016 Workshop (2016)
13. Duarte, L.A., Penatti, O.A.B., Almeida, J.: Bag of genres for video retrieval. In: SIBGRAPI. pp. 257–264 (2016)
14. Freund, Y., Iyer, R.D., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. J. Machine Learning Research 4, 933–969 (2003)
15. Jiang, Y.G., Dai, Q., Mei, T., Rui, Y., Chang, S.F.: Super fast event recognition in internet videos. IEEE Trans. Multimedia 17(8), 1174–1186 (2015)
16. Jiang, Y.G., Wang, Y., Feng, R., Xue, X., Zheng, Y., Yang, H.: Understanding and predicting interestingness of videos. In: AAAI. pp. 1113–1119 (2013)
17. Joachims, T.: Training linear svms in linear time. In: SIGKDD. pp. 217–226 (2006)
18. Kankanhalli, M.S., Lim, J.H.: Perspectives on Content-Based Multimedia Systems. Springer-Verlag, Inc., Secaucus, NJ, USA (2000)
19. Kludas, J., Bruno, E., Marchand-Maillet, S.: Adaptive multimedial retrieval: Retrieval, user, and semantics. chap. Information Fusion in Multimedia Information Retrieval, pp. 147–159 (2008)
20. Kokar, M.M., Tomasik, J.A., Weyman, J.: Formalizing classes of information fusion systems. Information Fusion 5(3), 189–202 (2004)
21. Kozorovitsky, A.K., Kurland, O.: Cluster-based fusion of retrieved lists. In: SIGIR. pp. 893–902 (2011)
22. Lam, V., Do, T., Phan, S., Le, D.D., Satoh, S., Duong, D.A.: NII-UIT at Mediaeval 2016 Predicting Media Interestingness task. In: Proc. of the MediaEval 2016 Workshop (2016)
23. Li, L.T., Pedronette, D.C.G., Almeida, J., Penatti, O.A.B., Calumby, R.T., Torres, R.S.: A rank aggregation framework for video multimodal geocoding. Multimedia Tools and Applications 73(3), 1323–1359 (2014)
24. Pedronette, D.C.G., Torres, R.S.: Image re-ranking and rank aggregation based on similarity of ranked lists. Pattern Recognition 46(8), 2350–2360 (2013)
25. Penatti, O.A.B., Li, L.T., Almeida, J., da S. Torres, R.: A visual approach for video geocoding using bag-of-scenes. In: ICMR. pp. 1–8 (2012)
26. Rayatdoost, S., Soleymani, M.: Ranking images and videos on visual interestingness by visual sentiment features. In: Proc. of the MediaEval 2016 Workshop (2016)
27. Shen, Y., Demarty, C.H., Duong, N.Q.K.: Technicolor at Mediaeval 2016 Predicting Media Interestingness task. In: Proc. of the MediaEval 2016 Workshop (2016)
28. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV. pp. 1470–1477 (2003)
29. Xu, B., Fu, Y., Jiang, Y.G.: BigVid at Mediaeval 2016: Predicting Interestingness in Images and Videos. In: Proc. of the MediaEval 2016 Workshop (2016)
30. Young, H.P.: An axiomatization of borda's rule. J. Economic Theory 9(1), 43–52 (1974)