Content-Based Multimedia Recommendation Systems: Definition and Application Domains

Yashar Deldjoo[1], Markus Schedl[2], Paolo Cremonesi[1], and Gabriella Pasi[3]

[1] Politecnico di Milano, Italy deldjooy@acm.org, paolo.cremonesi@polimi.it
[2] Johannes Kepler University Linz, Austria markus.schedl@jku.at
[3] University of Milano-Bicocca, Italy pasi@disco.unimib.it

Abstract. The goal of this work is to formally provide a general definition of a multimedia recommendation system (MMRS), in particular a content-based MMRS (CB-MMRS), and to shed light on different applications of multimedia content for solving a variety of tasks related to recommendation. We would like to disambiguate the fact that multimedia recommendation is not only about recommending a particular media type (*e.g.,* music, video), rather there exists a variety of other applications in which the analysis of multimedia input can be usefully exploited to provide recommendations of various kinds of information.

1 Introduction

The World Wide Web is a huge resource of digital multimedia information. In early years of the WWW, the available digital resources were mainly constituted of texts. For this reason, the first search engines and, later, content-based recommender systems relied merely on text analysis. Nowadays, the information available on the Web is provided by several different media types, which include text, audio, video, and images. Moreover, different media types can co-exist in documents such as for example Web pages. Vertical search engines and recommender systems have been developed to cope with the problem of accessing or recommending specific media objects. While some media types are not related to others (*e.g.,* texts), other media types, such as videos, can be considered as structured entities, possibly composite of other media types; for example a movie is a video object composed of a sequence of images and of an audio stream, and can further possibly carry a text (subtitles). The aim of this paper is twofold: on the one hand, we propose a general definition of content-based multimedia recommender system (CB-MMRS), which comprises both systems working with one media type (vertical approach) and systems working with multiple media types (*e.g.,* videos when exploiting the composite of image, audio, and textual information). Moreover we propose a general recommendation model of composite media objects, where the recommendation relies on the computation of distinct utility values, one for each media type in the composite object, and a final utility is computed by aggregating such values. This can pave the way for

IIR 2018, May 28-30, 2018, Rome, Italy.

novel recommendation techniques. As a second contribution, we discuss a variety of tasks where MM content can be exploited for effective recommendation, and we categorize them along different axes.

2 Content-Based Multimedia Recommendation Systems

We characterize a *content-based multimedia recommendation system* (CB-MMRS) by specifying its main components.

1. **Multimedia Items:** In the literature [1], a *multimedia item* (aka multimedia object or document) refers to an item which can be a text, image, audio, or video. Formally, a multimedia item I in its most general form is represented by the triple: $I = (I_V, I_A, I_T)$ in which I_V, I_A, I_T refer to the *visual*, *aural*, and *textual* components (aka modalities), respectively. While text can be considered an *atomic* media type (meaning it consists of a single textual modality I_T), other media types including audio, image, and video can be either *atomic* or *composite*, as in the latter case they may contain multiple modalities. For example, an audio item that represents a performance of a classical music piece can be seen as an atomic media type (using I_A). On the other hand, a pop song with lyrics can be regarded as composite (using I_A and I_T); an image of a scene is atomic (using I_V) while an image of a news article is composite (using I_V and I_T); finally a silent movie is atomic (using I_V) while a movie with sound is composite (using I_A, I_V and/or I_T). We still use the term multimedia item while referring to all these media types regardless of the fact that they are atomic or composite. A CB-MMRS is a system that is able to store and manage MM items.
2. **Multimedia Content-Based Representation:** Developing a CB-MMRS relies on content-based (CB) descriptions according to distinct modalities (I_V, I_A, I_T). These CB descriptors are usually extracted by applying some form of signal processing specific to each modality, and are described based on specific features. Examples of such features for images are color and texture; for text they include words and n-grams.
   A CB-MMRS is a system that is able to process MM items and represent each modality in terms of a feature vector $\mathbf{f_m} = [f_1\ f_2\ ...\ f_{n_m}] \in \mathbb{R}^{n_m}$ where $m \in \{V, A, T\}$ represents the visual, audio or textual modality.[4]
3. **Recommendation Model:** A CB-MMRS adopts a (personalized) recommendation model and provides suggestions for items by measuring the interest of user on CB characteristics of items in hand [2].
   Given a target user u, to whom the recommendation will be provided, and a collection of MM items $\mathcal{I} = \{I_1, I_2, ..., I_{|\mathcal{I}|}\}$, the task of MM recommendation is to identify the MM item i^* that satisfies

$$i^* = \underset{i}{\text{argmax}}\ \mathcal{R}(u, I_i),\ I_i\ \in\ \mathcal{I} \tag{1}$$

[4] We step aside our attention from end-to-end learning performed by deep neural networks where the intermediate step of feature extraction is not done explicitly.

Table 1: Conceptual goals, inputs, and outputs of MMRS. CF: Collaborative Filtering, CB: Content-based, pref= preference. MM-driven RS refer to systems which exploit multimedia content not necessarily to represent items or in the core recommendation model, exploiting MM content to represent users (*e.g.,* via analyzing their facial expressions). The output of MM-driven RS can be various form of information not necessarily bounded to MM such as RS exploiting MM content to recommend a non-media item *(e.g.,* recommend place of interest based on user-generated photos).

Approach	**Conceptual Goal**	**input**	**output**
CF-MMRS	Recommend me MM items by leveraging the preference of my peers/myself.	target user pref. + community pref.	MM item
CB-MMRS	Recommend me MM items based on the MM content of the items I liked in the past.	target user pref. + MM content	MM item
MM-driven RS	Give me recommendations based on the content of the MM items and other sources of information (ratings, context, *etc.*).	MM content + other info	various

where $\mathcal{R}(u, I_i)$ is the *estimated utility* of item i for the user u on the basis of which the items are ranked [3]. The utility can be only estimated by the RS to judge how much an item is *worth* being recommended, and its prediction lies at the core of a recommendation model. The utility estimation (or prediction) is done based on a particular *recommendation model*, *e.g.,* collaborative-filtering (CF) or content-based filtering (CBF), which typically involves knowledge about *users*, *items*, and the *core utility function* itself [2]. A comparison of such systems is provided in Table 1. While the community of RS for long has considered CF-MMRS or CB-MMRS using pure metadata (textual) as the only form of MMRS, in this paper, we focus our attention on CB-MMRS exploiting different media types and the constituting modalities. Depending on the number of modalities leveraged, we can categorize CB-MMRS as *unimodal* or *multimodal*. For example, unimodal CB-MMRS can produce satisfactory results for the recommendation of text (*e.g.,* a piece of news), but not for image, audio, and video. Users' diverse information needs are more likely to be satisfied by multimodal recommendation mechanisms. In a multimodal CB-MMRS, the estimated utility of item i to the user u can be decomposed into several specific utilities computed across each modality in a MM item:

$$\mathcal{R}(u, I_i) = F(\mathcal{R}_m(u, I_i)), \quad m \in \{V, A, T\} \tag{2}$$

where $\mathcal{R}_m(U, I_i)$ denotes the utility of item I_i for user u with regards to modality $m \in \{V, A, T\}$, and F is an *aggregation function* of the estimated utilities for each modality. Based on the semantics of the aggregation, different functions can be employed, each implying a particular interpretation of the affected process.

Aggregation operators can be roughly classified as *conjunctive*, *disjunctive*, and *averaging* [4,5]. Conjunctive operators include the minimum (min) and functions that are upper-bounded by the minimum

$$\mathcal{R}(u, I_i) <= \min\left(\mathcal{R}_m(u, I_i)\right), \quad \forall m \in \{V, A, T\} \tag{3}$$

Disjunctive operators include the maximum (max) and those functions lower-bounded by the maximum. Based on the choice of distinct aggregation operator, different aggregation values are obtained.
We make an illustrative example to clarify the importance of these aggregation operators. Suppose a MMRS should recommend a movie to a user. It is further known by the system that the user is likely to watch a fast-paced movie filmed with abrupt camera shot changes (visual), with rapid music tempo (audio), and with textual keywords that describe the movie as energetic or fast (textual). In such a case, if we set the aggregation function to (min), the system will follow a conservative/pessimistic approach and would require *all the three* audio plus visual plus textual modalities to contain the aforementioned properties, so the corresponding item can be considered as a good candidate for recommendation. Oppositely, the max operator adopts an optimistic approach and would only require *one of the three modalities* to include the desired property, making the utility of such item higher. Therefore, these two aggregation operators have distinct semantics which can be leveraged depending on the particular recommendation application at hand. The min and the max functions set a lower bound and an upper bound, respectively, for averaging aggregation operators, (*e.g.,* arithmetic mean, geometric mean, or harmonic mean). For instance, in the field of multimedia information retrieval (MMIR), it is common to use the weighted average linear combination

$$\mathcal{R}(u, I_i) = \sum_m w_m \mathcal{R}_m(u, I_i) \tag{4}$$

where w_m is a weight factor indicating the importance of modality m. When we focus our attention on a specific modality, the problem is similar to a standard CBF problem in which a linear model can be used among others, for example

$$\mathcal{R}_m(u, I_i) = \sum_j w_{mj} \mathcal{R}_{mj}(u, f_{mj}) \tag{5}$$

where w_{mj} is a weight factor indicating the importance of the feature f_{mj}, the j-th feature in modality m. Equations 4 and 5 are called the *inter-modality* and *intra-modality* fusion functions in MMIR. Application of different aggregation operators for CB-MMRS and generally MMRS remain open for exploitation in future works.

3 Multimedia Content for Tasks Related to Recommender Systems

Multimedia content can be leveraged in RS that recommend a media type or a non-media item to the user. Multimedia content can be also exploited for certain tasks that are related to RS, but are not directly part of the core recommendation approach or item model. Examples include the exploitation of web cam videos to identify the target user's emotional state [6], or in general her head/posture [7], which in turn can be used to personalize recommendations [8,9]. Another example is the use of audio content features to model transitions or learn sequences from music playlists, *e.g.,* continuously increasing energy level of songs in a playlist. Such information can then be used for automatic playlist generation or continuation [10]. The former example relates to the use of *multimedia* content in *context-aware recommender systems*, the latter to its use in *sequence recommendation*. We explore these different dimensions in the following.

3.1 Approaches that recommend multimedia items

The primary and foremost used application of multimedia content is constituted by MMRS *i.e.,* systems that recommend a particular media type to the user. In CB-MMRS, the media types constituting both the input and the output of the system are the same (*e.g.,* music recommendation based on music acoustic content plus target users' preferences); However in some applications the two media types can be also different, *e.g.,* recommending music for a given image with regards to the evoked emotions. We will explore these categories of recommendation in the following:

- **Audio recommendation:** As for audio recommendation, the most common application is *music recommendation* [10]. Examples of common audio features exploited in the music domain include energy, timbre, tempo, tonality, and more abstract ones based on deep learning [10].
- **Image recommendation:** In the image domain, some of the interesting examples include recommending *clothes* (in the fashion industry) and *paintings* (*e.g.,* in the tourism industry) among others. As for clothes recommendation, there exists a huge potential in the fashion industry, mainly for the economic value, to build personalized fashion RS. These systems can be built by taking into account metadata, reviews, previous purchasing patterns and visual appearance of products. Such recommendation can be done in two manners: (1) finding some pairs of objects that can be seen as *alternative* to a given image provided by the user (such as two pairs of jeans) and, (2) finding the ones which may be *complementary* (such recommending a pair of jeans matching a shirt). For example [11] proposed a CB-MMRS to provide personalized recommendation for a given clothes image by considering the visual appearances of clothes. The proposed system exploits visual features based on convolutional neural networks (pre-trained on 1.2M ImageNet images)

and uses a metric learning approach to find the visual similarity between a query image and the *complementary* items (second scenario). Some research works in the RS community [12] have criticized the above work in the sense that it treats the recommendation problem as a visual retrieval problem disregarding users' historical feedbacks on items as well as other factors beyond the visual dimension. The main novelty in [11], beside focusing on a novel clothes recommendation scenario is to examine the visual appearance of the items under investigation to overcome the 'cold start' problem.

– **Video recommendation** In the video domain, examples of target items include recommending movies, TV-series, movie clips, trailers, or user-generated content. In [13,14,15,16], the authors propose a video RS that exploits visual features complying with the mise-en-scene (stylistic aspect in a movie) and incorporate it in different CBF and CBF+CF systems to show that their proposed system can be replaced with similar systems using genre metadata and user-generated tags (in some cases). The authors show the possibility of utilizing such stylistic-based movie recommender systems in a real system [17] also for children [18] or combined with user's directly specified need in the form of a query by visual example [19].
A newer version of the authors' work [20] proposed advanced audio and visual descriptors originated from multimedia signal processing under a novel rank-aware hybridization approach to significantly improve quality of traditional RS over metadata.

3.2 Approaches that use multimedia items as input

Multimedia content can not only be used for a particular media-item recommendation (as illustrated above), but also there exists other applications where a MM item is used only as the input of such systems, while in the output another form of information is recommended. As listed in Table 1, we would like to call such system MM-driven RS to highlight that the output can be a non-media item. An example of such an application is provided below:

– **POI recommendation by analyzing user-generated photos:** [21] proposed a personalized travel recommendation system by leveraging the rich and freely available community-contributed photos and by considering demographical information such as gender, age and race in user profiles in order to provide effective personalized travel recommendation. The authors show that consideration of such attributes is effective for travel recommendation - especially providing a promising aspect for personalization. For this, the authors discuss the correlation between travel patterns and people characteristics by using information-theoretic measures.

3.3 Other approaches

An interesting but less-investigated area of research in CB-MMRS is recommending a piece of media (*e.g.,* music) based on its association with other media (*e.g.,* image) with regards to the evoked emotions, user-generated tags or other catalysts. For instance, [22] proposed a MMRS to automatically suggest music based on a set of images (paintings). The motivation is that the affective content of painting when harmonized with music can be effective for creating a fine art sideshow referred to as *emotion-based impressionism sideshow*. Emotion is used as the main enabler to find the association between the painting (input to the system) and the music (the output) which is done using Mixed Media Graph (MMG) model [23]. The proposed method uses a variety of visual features based on color, light and textures from the painting images as well as acoustic features such as melody, rhythm, tempo from the music where both categories of features are known to be affecting emotions.

Visual contextual advertisement is another very related application field of multimedia context in which the particular multimedia item currently being consumed by the user (*e.g.,* image or video) becomes the target for recommending advertisements. The goal here is to build a semantic match between two heterogeneous multimedia sources (*e.g.,* content of an image and the advertisement in textual form). [24] proposed a visual contextual advertisement system that suggests the most relevant advertisement for a given image without building a textual relation between the two heterogeneous sources (*i.e.,* it disregards the tags associated with images). The authors mention that there exists two main approaches for visual contextual advertisement: (1) based on image annotation, (2) based on feature translation model. In the first case, a model is trained based on a selection of labeled images which is leveraged to predict text on the test time given a new image. Manual labeling of the items is required which makes the approach prone to error or labor-intensive. The second approach builds a bridge between the two visual and textual feature spaces through a translation model and leveraging a language model to estimate the relevance of each advertisement *w.r.t* a given target image. In [24] the authors propose a knowledge-driven cross-media semantic matching framework that leverages two large high-quality knowledge sources ImageNet (for image) and Wikipedia (for text). The image-advertisement match is established in the respective knowledge sources.

3.4 Context-aware Recommendation

In context-aware or situation-aware recommender systems, which often enhance information about user–item interactions by considering time [25], user activity [26], or weather [27], among others, multimedia data can be user to create intermediate representations of items and match them with similar representations of context or users to effect recommendations. We exemplify this idea with the recent research topic of emotion-based matching, more precisely, using emotion information to match items to users and items to context entities.

Emotion-based matching of items and users: Here, the goal is to select items that match the target user's affective state. Eliciting the user's emotional state can be effected by requesting explicit feedback or by analyzing multimedia material, for instance, user-generated text [28], speech [29], or facial expressions in video [6], or a combination of audio and visual cues in video [30,31]. Likewise, describing items by affective terms can also be approached via content analysis. For instance, in the music domain, this task is commonly known as music emotion recognition (MER) [32]. Both tasks, i.e., inferring emotions from users and from items, come with their particular challenges, for instance, high variations in the intensity of users' facial expressions or subjectivity of perceived emotions when creating ground truth annotations of items. An even harder task, however, is to connect users with items in the affectively intended way. To do so, knowing about the target user's intent is crucial.

In the music domain, three fundamental intents or purposes of music listening have been identified [33]: self-awareness (*e.g.,* stimulating a reflection of people on their identity), social relatedness (*e.g.,* feeling closeness to friends and expressing identity), and arousal and mood regulation (*e.g.,* managing emotions). Several studies found that affect regulation is the most important purpose why people listen to music [33,34]. However, in which ways music preferences vary as a function of a listener's emotion, listening intent, and affective impact of listening to a certain emotionally laden music piece is still not well understood, and is further influenced by other psychological aspects such as the listener's personality [35].

Emotion-based matching of items and context entities: This task entails the establishment of relationships between items and contextual aspects. Affective information for items can again be elicited by multimedia analysis, those of contextual entities — in this scenario most commonly location [36] or weather [37] — by explicit user annotations. The recommender system then regards the emotion assigned to the contextual entity as proxy for the user's emotion, and matches items and users correspondingly.

To give an example, the system proposed in [36] recommends music pieces for locations (places of interest such as monuments). It uses emotions as intermediate representations of both. Based on online questionnaires, a limited set of places of interest are assigned affective labels. So are music pieces. Since the amount of potentially suited music pieces is, however, much larger than the number of interesting locations, an audio content-based auto-tagger for music [38] is trained on a small set of annotated pieces, and is subsequently used to predict the emotion tags for the remaining, unlabeled pieces. Recommendations for a target user at a given location are then made by ranking all music pieces according to the overlap (Jaccard coefficient) between their location and the location's affective labels.

3.5 Sequence Recommendation

In certain domains, recommendation of coherent or meaningful item sequences is preferred over recommendation of unordered item sets [39]. Examples include recommending online courses or exercises for e-learning, video clips in media streaming services, and automatic music playlist generation or continuation.

Recommending sequences of music pieces, i.e., playlists, is special for several reasons, most importantly the typically short duration and consumption time, the likely preference for repeated item consumption, and the strong emotional impact of music (cf. Section 3.4).

Approaches to automatic playlist generation or continuation can either learn directly from the sequences of items used for training, for instance, via sequential pattern mining [40], Markov models [41], or recurrent neural networks [42,43]. Alternatively or additionally, such approaches can also take content features into account. In the music domain, these descriptors may include tempo (beats per minute), timbre, or rhythm patterns and can be extracted through audio processing techniques. Other features relevant to describe music can be extracted from images like album covers [44] or video clips [45], which renders the task a multimedia content analysis problem. Using these content descriptors, playlists can either be created by computing similarities between songs, albums, or artists, or by defining constraints and creating the playlist in a way that fulfills these (as much as possible). The former approach aims at building coherent playlists in which consecutive tracks sound are as similar as possible, *e.g.,* [46,47]. The latter approach allows to define target characteristics, such as increasing tempo, high diversity of artists, or fixed start and end song [48,49].

Additionally, combining sequence recommendation with context-aware recommendation (cf. Section 3.4), playlists can be created based on hybrid methods that integrate the context of the listener and content-based similarity [50,51].

4 Conclusion and Future Work

In this work, we proposed a general definition of content-based multimedia recommender system (CB-MMRS). Moreover, we proposed a general recommendation model of composite media objects, where the recommendation relies on the computation of distinct utility values (for each media object) and a final utility is computed by aggregating such values. Finally, we presented variety of different applications where MM content is used not only as the target product, rather analyzed in the input of the system or to model the user to provide recommendation of various kinds of information.

References

1. Elleithy, K.: Advanced techniques in computing sciences and software engineering. Springer Science & Business Media (2010)
2. Ricci, F., Rokach, L., Shapira, B.: Recommender Systems: Introduction and Challenges. In: Recommender Systems Handbook. Springer (2015) 1–34

3. Aggarwal, C.C.: An introduction to recommender systems. In: Recommender Systems. Springer (2016) 1–28
4. Marrara, S., Pasi, G., Viviani, M.: Aggregation operators in information retrieval. Fuzzy Sets and Systems **324** (2017) 3–19
5. Tzeng, G.H., Huang, J.J.: Multiple attribute decision making: methods and applications. CRC press (2011)
6. Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent neural networks for emotion recognition in video. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. ICMI '15, New York, NY, USA, ACM (2015) 467–474
7. Deldjoo, Y., Atani, R.E.: A low-cost infrared-optical head tracking solution for virtual 3d audio environment using the nintendo wii-remote. Entertainment Computing **12** (2016) 9–27
8. Andjelkovic, I., Parra, D., O'Donovan, J.: Moodplay: Interactive mood-based music discovery and recommendation. In: Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization. UMAP '16, New York, NY, USA, ACM (2016) 275–279
9. Tkalčič, M., Burnik, U., Odić, A., Košir, A., Tasič, J.: Emotion-aware recommender systems – a framework and a case study. In Markovski, S., Gusev, M., eds.: ICT Innovations 2012, Berlin, Heidelberg, Springer Berlin Heidelberg (2013) 141–150
10. Schedl, M., Zamani, H., Chen, C.W., Deldjoo, Y., Elahi, M.: Current challenges and visions in music recommender systems research. International Journal of Multimedia Information Retrieval (2018) 1–22
11. McAuley, J., Targett, C., Shi, Q., Van Den Hengel, A.: Image-based recommendations on styles and substitutes. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM (2015) 43–52
12. He, R., McAuley, J.: Vbpr: Visual bayesian personalized ranking from implicit feedback. (2016)
13. Deldjoo, Y., Elahi, M., Cremonesi, P., Garzotto, F., Piazzolla, P., Quadrana, M.: Content-based video recommendation system based on stylistic visual features. Journal on Data Semantics **5**(2) (2016) 99–113
14. Deldjoo, Y., Cremonesi, P., Schedl, M., Quadrana, M.: The effect of different video summarization models on the quality of video recommendation based on low-level visual. In: Content-Based Multimedia Indexing (CBMI), 2017 15th International Workshop on. ACM. (2017)
15. Deldjoo, Y., Elahi, M., Cremonesi, P., Garzotto, F., Piazzolla, P.: Recommending movies based on mise-en-scene design. In: Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems, ACM (2016) 1540–1547
16. Deldjoo, Y., Elahi, M., Quadrana, M., Cremonesi, P.: Using visual features based on mpeg-7 and deep learning for movie recommendation. International Journal of Multimedia Information Retrieval (2018)
17. Elahi, M., Deldjoo, Y., Bakhshandegan Moghaddam, F., Cella, L., Cereda, S., Cremonesi, P.: Exploring the semantic gap for movie recommendations. In: Proceedings of the Eleventh ACM Conference on Recommender Systems, ACM (2017) 326–330
18. Deldjoo, Y., Frà, C., Valla, M., Paladini, A., Anghileri, D., Tuncil, M.A., Garzotta, F., Cremonesi, P., et al.: Enhancing childrens experience with recommendation systems. In: Workshop on Children and Recommender Systems (KidRec'17)-11th ACM Conference of Recommender Systems. (2017) N–A

19. Deldjoo, Y., Fra, C., Valla, M., Cremonesi, P.: Letting users assist what to watch: An interactive query-by-example movie recommendation system. (2017)
20. Deldjoo, Y., Constantin, M.G., Schedl, M., Ionescu, B., Cremonesi, P.: Mmtf-14k: A multifaceted movie trailer feature dataset for recommendation and retrieval. In: Proceedings of the 9th ACM Multimedia Systems Conference, ACM (2018)
21. Cheng, A.J., Chen, Y.Y., Huang, Y.T., Hsu, W.H., Liao, H.Y.M.: Personalized travel recommendation by mining people attributes from community-contributed photos. In: Proceedings of the 19th ACM international conference on Multimedia, ACM (2011) 83–92
22. Li, C.T., Shan, M.K.: Emotion-based impressionism slideshow with automatic music accompaniment. In: Proceedings of the 15th ACM international conference on Multimedia, ACM (2007) 839–842
23. Pan, J.Y., Yang, H.J., Faloutsos, C., Duygulu, P.: Automatic multimedia cross-modal correlation discovery. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2004) 653–658
24. Zhang, W., Tian, L., Sun, X., Wang, H., Yu, Y.: A semantic approach to recommending text advertisements for images. In: Proceedings of the sixth ACM conference on Recommender systems, ACM (2012) 179–186
25. Herrera, P., Resa, Z., Sordo, M.: Rocking around the clock eight days a week: an exploration of temporal patterns of music listening. In: Proceedings of the ACM Conference on Recommender Systems: Workshop on Music Recommendation and Discovery (WOMRAD 2010). (2010) 7–10
26. Wang, X., Rosenblum, D., Wang, Y.: Context-aware Mobile Music Recommendation for Daily Activities. In: Proceedings of the 20^{th} ACM International Conference on Multimedia, Nara, Japan, ACM (2012) 99–108
27. Pettijohn, T., Williams, G., Carter, T.: Music for the seasons: Seasonal music preferences in college students. Current Psychology (2010) 1–18
28. Dey, L., Asad, M.U., Afroz, N., Nath, R.P.D.: Emotion extraction from real time chat messenger. In: 2014 International Conference on Informatics, Electronics Vision (ICIEV). (May 2014) 1–5
29. Erdal, M., Kächele, M., Schwenker, F. In: Emotion Recognition in Speech with Deep Learning Architectures. Springer International Publishing, Cham (2016) 298–311
30. Kaya, H., G F.: Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image and Vision Computing **65** (2017) 66–75 Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.
31. Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., Anbarjafari, G.: Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing (2017) 1–1
32. Yang, Y.H., Chen, H.H.: Machine recognition of music emotion: A review. Transactions on Intelligent Systems and Technology **3**(3) (May 2013)
33. Schäfer, T., Sedlmeier, P., Stdtler, C., Huron, D.: The psychological functions of music listening. Frontiers in Psychology **4**(511) (2013) 1–34
34. Lonsdale, A.J., North, A.C.: Why do we listen to music? A uses and gratifications analysis. British Journal of Psychology **102**(1) (February 2011) 108–134
35. Ferwerda, B., Schedl, M., Tkalčič, M.: Personality & Emotional States: Understanding Users Music Listening Needs. In: Extended Proceedings of the 23rd International Conference on User Modeling, Adaptation and Personalization (UMAP 2015), Dublin, Ireland (June–July 2015)

36. Kaminskas, M., Ricci, F., Schedl, M.: Location-aware Music Recommendation Using Auto-Tagging and Hybrid Matching. In: Proceedings of the 7th ACM Conference on Recommender Systems (RecSys), Hong Kong, China (October 2013)
37. Coviello, L., Sohn, Y., Kramer, A.D.I., Marlow, C., Franceschetti, M., Christakis, N.A., Fowler, J.H.: Detecting emotional contagion in massive social networks. PLOS ONE **9**(3) (03 2014) 1–6
38. Seyerlehner, K., Sonnleitner, R., Schedl, M., Hauger, D., Ionescu, B.: From Improved Auto-taggers to Improved Music Similarity Measures. In: Proceedings of the 10th International Workshop on Adaptive Multimedia Retrieval (AMR 2012), Copenhagen, Denmark (October 2012)
39. Quadrana, M., Cremonesi, P., Jannach, D.: Sequence-aware recommender systems. CoRR **abs/1802.08452** (2018)
40. Lu, E.H.C., Lin, Y.W., Ciou, J.B.: Mining mobile application sequential patterns for usage prediction. In: 2014 IEEE International Conference on Granular Computing (GrC). (Oct 2014) 185–190
41. Chen, S., Moore, J.L., Turnbull, D., Joachims, T.: Playlist prediction via metric embedding. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM (2012) 714–722
42. Hidasi, B., Karatzoglou, A., Baltrunas, L., Tikk, D.: Session-based recommendations with recurrent neural networks. CoRR **abs/1511.06939** (2015)
43. Vall, A., Quadrana, M., Schedl, M., Widmer, G., Cremonesi, P.: The Importance of Song Context in Music Playlists. In: Proceedings of the Poster Track of the 11th ACM Conference on Recommender Systems (RecSys), Como, Italy (2017)
44. Lībeks, J., Turnbull, D.: You can judge an artist by an album cover: Using images for music annotation. IEEE MultiMedia **18**(4) (April 2011) 30–37
45. Schindler, A., Rauber, A.: Harnessing music-related visual stereotypes for music information retrieval. ACM Trans. Intell. Syst. Technol. **8**(2) (October 2016) 20:1–20:21
46. Pohle, T., Knees, P., Schedl, M., Pampalk, E., Widmer, G.: "Reinventing the Wheel": A Novel Approach to Music Player Interfaces. IEEE Transactions on Multimedia **9** (2007) 567–575
47. Knees, P., Pohle, T., Schedl, M., Widmer, G.: Combining Audio-based Similarity with Web-based Data to Accelerate Automatic Music Playlist Generation. In: Proceedings of the 8th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR), Santa Barbara, CA, USA (2006)
48. Chen, S., Moore, J.L., Turnbull, D., Joachims, T.: Playlist prediction via metric embedding. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '12, New York, NY, USA, ACM (2012) 714–722
49. Flexer, A., Schnitzer, D., Gasser, M., Widmer, G.: Playlist Generation Using Start and End Songs. In: Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR), Philadelphia, PA, USA (2008)
50. Cheng, Z., Shen, J.: Just-for-Me: An Adaptive Personalization System for Location-Aware Social Music Recommendation. In: Proceedings of the 4th ACM International Conference on Multimedia Retrieval (ICMR), Glasgow, UK (2014)
51. Reynolds, G., Barry, D., Burke, T., Coyle, E.: Towards a Personal Automatic Music Playlist Generation Algorithm: The Need for Contextual Information. In: Proceedings of the 2nd International Audio Mostly Conference: Interaction with Sound, Ilmenau, Germany (2007) 84–89