J Intell Inf Syst (2013) 41:523–539
DOI 10.1007/s10844-013-0247-6

The neglected user in music information retrieval research

Markus Schedl · Arthur Flexer · Julián Urbano

Received: 14 November 2012 / Revised: 13 March 2013 / Accepted: 9 May 2013 /
Published online: 18 July 2013

Abstract Personalization and context-awareness are highly important topics in research on Intelligent Information Systems. In the fields of Music Information Retrieval (MIR) and Music Recommendation in particular, user-centric algorithms should ideally provide music that perfectly fits each individual listener in each imaginable situation and for each of her information or entertainment needs. Even though preliminary steps towards such systems have recently been presented at the "International Society for Music Information Retrieval Conference" (ISMIR) and at similar venues, this vision is still far away from becoming a reality. In this article, we investigate and discuss literature on the topic of user-centric music retrieval and reflect on why the breakthrough in this field has not been achieved yet. Given the different expertises of the authors, we shed light on why this topic is a particularly challenging one, taking *computer science* and *psychology* points of view. Whereas the computer science aspect centers on the problems of user modeling, machine learning,

This research is supported by the Austrian Science Fund (FWF): P22856, P25655, and P24095, by the Spanish Government: HAR2011-27540, and by the European Commission, FP7 (Seventh Framework Programme), ICT-2011.1.5 Networked Media and Search Systems, grant agreement no. 287711. The research leading to these results has received funding from the European Union Seventh Framework Programme FP7 / 2007–2013 through the PHENICX project under grant agreement no. 601166.

M. Schedl (✉)
Department of Computational Perception, Johannes Kepler University, Linz, Austria
e-mail: markus.schedl@jku.at

A. Flexer
Austrian Research Institute for Artificial Intelligence, Vienna, Austria
e-mail: arthur.flexer@ofai.at

J. Urbano
Department of Computer Science, University Carlos III, Leganés, Spain
e-mail: jurbano@inf.uc3m.es

and evaluation, the psychological discussion is mainly concerned with proper experimental design and interpretation of the results of an experiment. We further present our ideas on aspects crucial to consider when elaborating user-aware music retrieval systems.

Keywords User-centric music retrieval · Experimental design · Evaluation · Interpretation

1 Why care about the user?

In our discussion of the importance and the challenges of development and evaluation in Music Information Retrieval (MIR) we distinguish between *systems-based* and *user-centric* MIR. We define systems-based MIR as all research concerned with laboratory experiments existing solely in a computer, e.g. evaluation of algorithms on digital databases. In contrast, user-centric MIR always involves human subjects and their interaction with MIR systems.

Systems-based MIR has traditionally focused on computational models to describe universal aspects of human music perception, for instance, via elaborating musical feature extractors or similarity measures. Doing so, the existence of an objective "ground truth" is assumed, against which corresponding music retrieval algorithms (e.g., playlist generators or music recommendation systems) are evaluated. To give a common example, music retrieval approaches have been evaluated via genre classification experiments for years. Although it was shown already in 2003 that musical genre is an ill-defined concept (Aucouturier and Pachet 2003), genre information still serves as a proxy to vaguely assess music similarity and retrieval approaches in systems-based MIR.

On the way towards user-centric MIR, the coarse and ambiguous concept of genre should either be treated in a personalized way or replaced by the concept of similarity. When humans are asked to judge the similarity between two pieces of music, however, certain other challenges need to be faced. Common evaluation strategies typically do not take into account the musical expertise and taste of the users. A clear definition of "similarity" is often missing too. It might hence easily occur that two users apply a very different, individual notion of similarity when assessing the output of music retrieval systems. While a first person may experience two songs as rather dissimilar due to very different lyrics, a second one may feel a much higher resemblance of the very same songs because of a similar instrumentation. Similarly, a fan of Heavy Metal music might perceive a Viking Metal track as dissimilar to a Death Metal piece, while for the majority of people the two will sound alike. Scientific evidence for this subjective perception of musical similarity can be found, for instance, in Law and von Ahn (2009) in which a new kind of "game with a purpose" is proposed. Named "TagATune", the aim of this 2-player-game is to decide if two pieces the players listen to simultaneously are the same or not. To this end, they are allowed to exchange free-form labels, tags, or other text. In a bonus round, players are presented three songs, one seed and two target songs. They now have to decide, which of the two targets is more similar to the seed. Based on an analysis of the data collected in this bonus round, Wolff and Weyde (2011) show that there are many tuples on which users do not agree. A more decent investigation of perceptual human similarity

is performed in Novello et al. (2006), where Novello et al. analyze concordance of relative human similarity judgments gathered by an experiment similar to the TagATune bonus rounds. The experiment included 36 participants who had to judge the same set of 102 triads each. Although the authors report statistically significant concordance values for 95 % of the triads (measured via Kendall's coefficient of rank correlation), only about half of the triads show a correlation value higher than 0.5, which is frequently taken as indicator of a moderate correlation.

Analyzing how users organize their music collections and which methods they apply to browse them or seek for particular music has not been of major interest in the MIR community, although this topic is certainly related to user-centric MIR. Work in the corresponding research area is carried out to a large extent by Sally Jo Cunningham and colleagues, who dubbed it "music information behavior". For instance, Cunningham et al. (2004) reports on a study performed via interviews and on-site observations, aiming at investigating how people organize their music collections. Their findings include that (i) a person's physical music collection is frequently divided into "active items" and "archival items", (ii) albums are frequently sorted according to date of purchase, release date, artist in alphabetic order, genre, country of origin, most favorite to least favorite, or recency of being played, and (iii) music is frequently organized according to the intended use, for instance, a particular event or occasion.

Looking into user behavior when it comes to constructing playlists, Cunningham et al. (2006) carried out a qualitative study based on user questionnaires and postings of related web sites. They found that users frequently start creating a playlist by browsing through their music collections in a linear manner or by considering their recent favorite songs. Cunningham et al. further criticize that most music retrieval systems are missing a function to explicitly exclude songs with a particular attribute (e.g., music of a particular genre or by a certain artist). Given the results of another study (Cunningham et al. 2005), which aimed at assessing which songs are the most hated ones, such a function would be vital, though.

More recent work looks into music listening and organization behavior via online surveys (Kamalzadeh et al. 2012) or tackle specific user groups, for instance, homeless people in North America (Woelfer and Lee 2012). The former study found that the most important attributes used to organize music are artist, album, and genre. When it comes to creating playlists, also mood plays an important role. Furthermore, the study showed a strong correlation between user activities (in particular, high attention and low attention activities) and aspects such as importance, familiarity, and mood of songs, as well as willingness to interact with the player. The latter study (Woelfer and Lee 2012) investigates the reasons for listening to music, among homeless people. It reveals that calming down, help to get through difficult times, and just to relieve boredom are the most important driving factors why homeless young people in Vancouver, Canada, listen to music.

The above examples and analyses illustrate that there are many aspects that influence what a human perceives as similar in a musical context. According to Schedl and Knees (2011), these aspects can be grouped into three different categories: *music content*, *music context*, and *user context*. Here we extend our previous categorization (Schedl and Knees 2011) by a fourth set of aspects, the *user properties*. Examples for each category are given in Fig. 1. Broadly speaking, *music content* refers to all aspects that are encoded in and can be inferred from the audio signal, while *music*

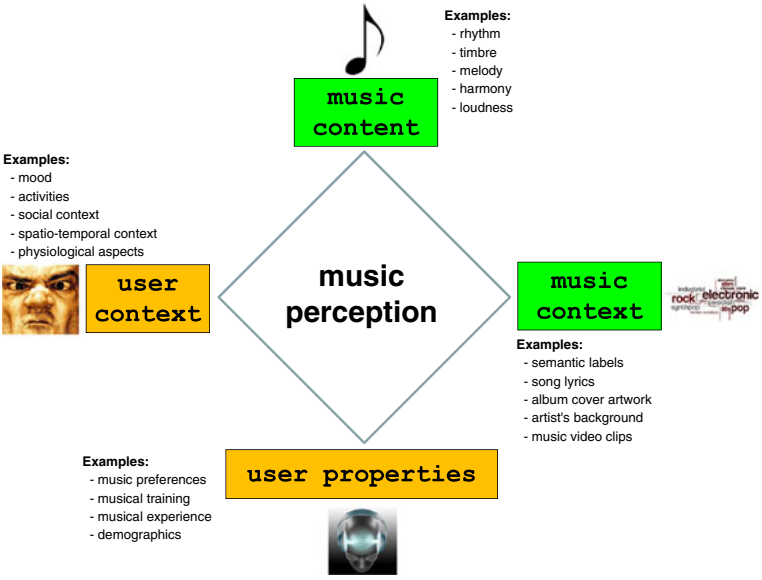

Fig. 1 Factors that influence human music perception

context includes factors that cannot be extracted directly from the audio, but are nevertheless related to the music item, artist, or performer. For instance, rhythmic structure, melody, and timbre features belong to the former category, whereas information about the artist's cultural or political background, collaborative semantic labels, and album cover artwork fall into the latter category. While *user context* aspects represent dynamic and frequently changing factors, such as the user's current social context or activity, *user properties* refer to constant or only slowly changing features of the user, such as her music taste or skills in playing instruments. The incorporation of user context and user properties into our model of music perception is also justified by the analysis reported in Hargreaves et al. (2005) about how people communicate using music. In particular, Hargreaves et al. highlight the importance of "non-music context" both for communicating through music and for the listeners' perception of music. The authors give some examples of such context categories and particular aspects: social and cultural context (political and national context), everyday situations (work, leisure, consumer, entertainment), presence/absence of others (live, audience, recorder).

It is exactly this multifaceted and individual way of music perception that has largely been neglected so far when elaborating and evaluating music retrieval approaches, but should be given more attention, in particular considering the trend towards personalized and context-aware systems (Liem et al. 2011; Schedl and Knees 2011).

A *personalized system* is one that incorporates information about the user into its data processing part (e.g., a particular user taste for a movie genre). A *context-aware system*, in contrast, takes into account dynamic aspects of the user context when processing the data (e.g., location and time where/when a user issues a query). Although the border between personalization and context-awareness may appear fuzzy from this definition, in summary, personalization usually refers to the incorporation

of more static, general user preferences, whereas context-awareness refers to the fact that frequently changing aspects of the user's environmental, psychological, and physiological context are considered. Given our categorization of aspects influencing music perception (Fig. 1), generally speaking, personalization draws on factors in the category *user properties*, whereas context-aware models encompass aspects of the *user context*.

In discussing these aspects of user-centric MIR, we will take both a computer science and psychological point of view. The computer science aspect is mainly concerned with the algorithmic and computational challenges of modeling individual or groups of users in MIR. Our psychological approach concentrates on proper experimental design and interpretation of results. Of course we are aware that psychology is a much broader field and that music psychology in particular tries to explain both musical behavior and musical experience as a whole with psychological methods. Discussion of this broader field of common interests is beyond the scope of this paper and we like to point interested readers to a joint presentation of an eminent MIR researcher and a psychologist elaborating on the sometimes complicated dialog of the two disciplines at last year's major conference in the MIR field (ISMIR 2012[1]) (Aucouturier and Bigand 2012).

The remainder of this paper is organized as follows. Section 2 reviews approaches that, in one way or the other, take the user into account when building music retrieval systems. We also discuss here the role of the user in communities other than MIR and analyze what the MIR community can learn from others. Evaluation strategies for investigating user-centric MIR are discussed in Section 3. In Section 4, we eventually summarize important factors when creating and evaluating user-aware music retrieval systems.

2 How to model the user?

Existing personalized and user-aware systems typically model the user in a very simplistic way. For instance, it is common in *collaborative filtering* approaches (Linden et al. 2003; Sarwar et al. 2001) to build user profiles only from information about a user u expressing an interest in item i. As an indicator of interest may serve, for example, a click on a particular item, a purchasing transaction, or in MIR the act of listening to a certain music piece. Such indications, in their simplest form, are stored in a binary matrix where element $r(u, i)$ denotes the presence or absence of a connection between user u and item i. In common recommendation systems, a more fine-grained scale for modeling the user interest in an item is typically employed—users frequently rate items according to a Likert-type scale, e.g., by assigning one to five stars to it. Matrix factorization techniques are subsequently applied to recommend novel items (Koren et al. 2009).

In the following, we first analyze the role of the user in literature about MIR (Section 2.1). We then look at how other communities, in particular the Recommendation Systems community, address the user and what the MIR community can learn from these (Section 2.2).

[1] http://ismir2012.ismir.net/

2.1 What about the user in MIR?

Taking a closer look at literature about context-aware retrieval and recommendation in the music domain, we can see that approaches differ considerably in terms of how the user context is defined, gathered, and incorporated. A summary and categorization of corresponding scientific works can be found in Table 1. The majority of approaches rely solely on one or a few aspects (temporal features in Cebrián et al. 2010, listening history and weather conditions in Lee and Lee 2007, for instance), whereas more comprehensive user models are rare in MIR. One of the few exceptions is Cunningham et al.'s study (Cunningham et al. 2008) that investigates if and how various factors relate to music taste (e.g., human movement, emotional status, and external factors such as temperature and lightning conditions). Based on the findings, the authors present a fuzzy logic model to create playlists.

There further exists some work that assumes a mobile music consumption scenario. The corresponding systems frequently aim at matching music with the current pace of a walker or jogger, e.g. Moens et al. (2010), and Biehl et al. (2006). Such systems typically try to match the user's heartbeat with the music played (Liu et al. 2009). However, almost all proposed systems require additional hardware for context logging, e.g. Elliott and Tomlinson (2006), Dornbush et al. (2007), and Cunningham et al. (2008).

Table 1 Categorization of literature about music retrieval including user aspects

Features	music content	(Zhang et al. 2009), (Knees and Widmer 2007), (Nürnberger and Detyniecki 2003)
	music context	(Kaminskas and Ricci 2011), (Zhang et al. 2009), (Pohle et al. 2007), (Knees and Widmer 2007)
	user-centric	(Cebrián et al. 2010) – few features, (Lee and Lee 2007) – few features, (Cunningham et al. 2008) – many features, (Xue et al. 2009) – features at different levels
Personalization	relevance feedback	(Knees and Widmer 2007)
	user-adjustable weights	(Zhang et al. 2009), (Pohle et al. 2007), (Nürnberger and Detyniecki 2003)
Context-Awareness	restricted to "sports"	(Moens et al. 2010), (Liu et al. 2009), (Biehl et al. 2006), (Elliott and Tomlinson 2006), (Dornbush et al. 2007), (Cunningham et al. 2008)
	restricted to "driving a car"	(Baltrunas et al. 2011)
	restricted to "places of interest"	(Kaminskas and Ricci 2011)
Evaluation	no user involvement or not reported	(Cebrián et al. 2010), (Pohle et al. 2007), (Nürnberger and Detyniecki 2003)
	precompiled user-generated data sets	(Xue et al. 2009), (Knees et al. 2007), (Lee and Lee 2007)
	user response to single question	(Kaminskas and Ricci 2011), (Liu et al. 2009), (Moens et al. 2010), (Biehl et al. 2006)
	multifaceted questionnaire	(Bogdanov and Herrera 2011), (Firan et al. 2007)

In Kaminskas and Ricci (2011) a system that matches tags describing a particular place with tags describing music is presented. Employing text-based similarity measures between the multimodal sets of tags, Kaminskas and Ricci propose their system for location-based music recommendation. Baltrunas et al. (2011) suggest a context-aware music recommender system for music consumption while driving. Although the authors take into account eight different contextual factors (e.g., driving style, mood, road type, weather, traffic conditions), their application scenario is quite restricted and their system relies on explicit human feedback, which is burdensome.

Zhang et al. present *CompositeMap* (Zhang et al. 2009), a model that takes into account similarity aspects derived from music content as well as social factors. The authors propose a multimodal music similarity measure and show its applicability to the task of music retrieval. They also allow a simple kind of personalization of this model by letting the user weight the individual music dimensions on which similarity is estimated. However, they do neither take the user context into consideration, nor do they try to learn a user's preferences.

Pohle et al. (2007) present preliminary steps towards a simple personalized music retrieval system. Based on a clustering of community-based tags extracted from *Last.fm*, a small number of musical concepts are derived using *Non-Negative Matrix Factorization* (NMF) (Lee and Seung 1999; Xu et al. 2003). Each music artist or band is then described by a "concept vector". A user interface allows for adjusting the weights of the individual concepts, based on which artists that best match the resulting distribution of the concepts are recommended to the user. Zhang et al. (2009) propose a very similar kind of personalization strategy via user-adjusted weights.

Knees and Widmer (2007) present an approach that incorporates *relevance feedback* (Rocchio 1971) into a text-based music search engine (Knees et al. 2007) to adapt the retrieval process to user preferences. The search engine proposed by Knees et al. builds a model from music content features (MFCCs) and music context features (term vector representations of artist-related Web pages). To this end, a weight is computed for each (term, music item)-pair, based on the term vectors. These weights are then smoothed, taking into account the closest neighbors according to the content-based similarity measure (Kullback–Leibler divergence on Gaussian Mixture Models of the MFCCs). To retrieve music via natural language queries, each textual query issued to the system is expanded via a *Google* search, resulting again in a term weight vector. This query vector is subsequently compared to the smoothed weight vectors describing the music pieces, and those with smallest distance to the query vector are returned.

Nürnberger and Detyniecki (2003) present a variant of the *Self-Organizing Map* (SOM) (Kohonen 2001) that is based on a model that adapts to *user feedback*. To this end, the user can move data items on the SOM. This information is fed back into the SOM's codebook, and the mapping is adapted accordingly.

Xue et al. (2009) present a *collaborative personalized search model* that alleviates the problems of *data sparseness* and *cold-start for new users* by combining information on different levels (individual, interest group, and global). Although not explicitly targeted at music retrieval, the idea of integrating data about the user, his peer group, and global data to build a social retrieval model might be worth considering for MIR purposes.

The problem with the vast majority of approaches presented so far is that evaluation is still carried out without sufficient user involvement. For instance, Cebrián

et al. (2010), Pohle et al. (2007), and Nürnberger and Detyniecki (2003) seemingly do not perform any kind of evaluation involving real users, or at least do not report it. Some approaches are evaluated on user-generated data, but do not request feedback from real users during the evaluation experiments. For example, Knees et al. (2007) make use of collaborative tags stored in a database to evaluate the proposed music search engine. Similarly, Lee and Lee (2007) rely on data sets of listening histories and weather conditions, and Xue et al. (2009) use a corpus of Web search data. Even if real users are questioned during evaluation, their individual properties (such as taste, expertise, or familiarity with the music items under investigation) are regularly neglected in evaluation experiments. In these cases, evaluation is typically performed to answer a very narrow question in a restricted setting. To give an example, the work on automatically selecting music while doing sports, e.g. Liu et al. (2009), Moens et al. (2010), and Biehl et al. (2006), is evaluated on the very question of whether pace or heartbeat of the user does synchronize with the tempo of the music. Likewise Kaminskas and Ricci's work on matching music with places of interest (Kaminskas and Ricci 2011), even though it is evaluated by involving real users, comprises only the single question of whether the music suggested by their algorithm is suited for particular places of interest or not. Different dimensions of the relation between images and music are not addressed. Although this is perfectly fine for the intended use cases, such highly specific evaluation settings are not able to provide answers to more general questions of music retrieval and recommendation, foremost because these settings fail at offering *explanations* for the (un)suitability of the musical items under investigation.

An evaluation approach that tries to alleviate this shortcoming is presented in Bogdanov and Herrera (2011), where subjective listening tests to assess music recommendation algorithms are conducted using a multifaceted questionnaire. Besides investigating the enjoyment a user feels when listening to the recommended track ("liking"), the authors also ask for the user's "listening intention", whether or not the user knows the artist and song ("familiarity"), and whether he or she would like to request more similar music ("give-me-more"). A similar evaluation scheme is suggested by Firan et al. (2007), though they only investigate liking and novelty.

In summary, almost all approaches reported are still more systems-based than user-centric.

2.2 What about the user in other communities?

Other research communities, in particular the Recommendation Systems (RS) and the Text-IR communities, include the user much more comprehensively in evaluation. An overview of relevant literature in these two areas is given below.

When looking at the RS community, there is a long tradition in using the systems-based performance measure of Root Mean Square Error (RMSE) to measure recommendation quality (Ricci et al. 2011). This measure is typically computed and investigated in leave-one-out experiments. However, a few years ago the RS community started to recognize the importance of user-centric evaluation strategies, and reacted accordingly. Pu et al. (2011) present a highly detailed user-centric evaluation framework, which make use of psychometric user satisfaction questionnaires. They analyze a broad variety of factors organized into four categories: perceived system qualities, user beliefs, user attitudes, and behavioral intentions. In particular, Pu and

Chen highlight (i) *perceived accuracy*, i.e. the degree to which users feel that the recommendations match their preferences, (ii) *familiarity*, i.e. whether users have previous knowledge about the recommended items, (iii) *novelty*, i.e. whether novel items are recommended, (iv) *attractiveness*, i.e. whether recommended items are capable of stimulating a positive emotion of interest or desire, (v) *enjoyability*, i.e. whether users have joyful experience with the suggested items, (vi) *diversity* of the recommended items, and (vii) *context compatibility*, i.e. whether the recommended items fit the current user context, such as the user's mood or activity. In addition to these aspects, Pu and Chen propose user questions that assess the *perceived usefulness* and *transparency* of a recommender, as well as *user intentions* towards the recommendation system.

A similar study, though not as comprehensive, is presented by Dooms et al. (2011). The authors use a questionnaire and ask users to explicitly rate different qualities of the recommender system under investigation using a Likert-type 5-point scale. In addition, they also look into implicit user feedback, analyzing the user interaction with the system. Based on this input, Dooms et al. identify eight relevant aspects that are important to users when interacting with recommendation systems: match between recommended items and user interests, familiarity of the recommended items, ability to discover new items, similarity between recommended items, explanation why particular items are recommended, overall satisfaction with the recommendations, trust in the recommender, and willingness to purchase some of the recommended items.

To further underline the importance RS researchers attribute to user-centric evaluation, a workshop series dedicated to the very topic of "User-centric Evaluation of Recommender Systems and Their Interface",[2] held in conjunction with the "ACM Conference on Recommender Systems", came into life in 2010 (Knijnenburg et al. 2010).

Some of the user-centric aspects addressed in RS literature can also be found in IR research. In particular, the properties of *novelty* and *diversity* (Clarke et al. 2008) as well as *transparency* (Tunkelang 2008), i.e. explaining why a particular item has been returned by a retrieval systems, are frequently mentioned. Also the aspect of *redundancy*, i.e. omitting redundant results that are annoying for most users, is addressed (Zhang et al. 2002).

The IR community is thus also seeing a paradigm shift in evaluation and performance measurement, away from the traditional systems-based relevance measures, such as precision, recall, precision at k retrieved documents (P@k), mean average precision (MAP), or discounted cumulative gain (DCG), e.g. Baeza-Yates and Ribeiro-Neto (2011), towards considering *user interaction* and system usage (Azzopardi et al. 2011). A vital role is hence played by emphasizing interaction with information, instead of passive user consumption of documents or items returned by a retrieval system (Järvelin 2012). Järvelin as well as Callan et al. (2007) propose a shift in the general design of IR systems, away from the concept of users finding documents, towards information interaction via clustering, linking, summarizing, arranging, and social networks.

[2] http://ucersti.ieis.tue.nl

3 How to evaluate user-centric MIR?

In what follows we will argue that whereas evaluation of systems-based MIR has quite matured, evaluation of user-centric MIR is still in its infancy.

3.1 Systems-based and user-centric MIR experiments

Let us start by reviewing what the nature of experiments is in the context of MIR. The basic structure of MIR experiments is the same as in any other experimental situation: the objective is to measure the effect of different treatments on a dependent variable. Typical dependent variables in systems-based MIR are various performance measures like accuracy, precision, root mean squared error or training time; and the treatments are the different algorithms to evaluate and compare, or different parametrizations of the same algorithm. A standard computer experiment is genre classification, where the treatments are different types of classification algorithm, say algorithms A and B, and the dependent variable is the achieved accuracy. But there are many other factors that might influence the results of the algorithms. For example, the musical expertise of the end user plays an important role in how good genre classification algorithms are perceived: as mentioned, a Heavy Metal fan is able to distinguish between Viking Metal and Death Metal, while most people do not. As another example, consider a fan of Eric Clapton that wishes to find similar music and a recommender system suggests Cream or Derek and the Dominos, which are bands surely known by this specific user but rather not by every general user. Any factor that is able to influence the dependent variables should be part of the experimental design, such as the musical expertise or known artists in the examples above. The important thing to note is that for systems-based MIR, which uses only computer experiments, it is comparably easy to control all important factors which could have an influence on the dependent variables. This is because the number of factors is both manageable and controllable, since the experiments are being conducted on computers and not in the real world. Indeed, the only changing factor is the algorithm to use.

Already early in the history of MIR research, gaps concerning the evaluation of MIR systems have been identified. Futrelle and Downie (2003), in their 2003 review of the first three years of the ISMIR conference, identify two major problems: (i) no commonly accepted means of comparing retrieval techniques, (ii) few, if any, attempts to study potential users of MIR systems. The first problem concerns the lack of standardized frameworks to evaluate computer experiments, while the second problem concerns the barely existing inclusion of users in MIR studies. Flexer (2006), in his review of the 2004 ISMIR conference (Buyoli and Loureiro 2004), argues for the necessity of statistical evaluation of MIR experiments. He presents minimum requirements concerning statistical evaluation by applying fundamental notions of statistical hypothesis testing to MIR research. But his discussion is concerned with systems-based MIR: the example used throughout the paper is that of automatic genre classification based on audio content analysis.

Statistical testing is needed to assess the confidence in that the observed effects on the dependent variables are caused by the varied independent variables and not by mere chance, i.e. to ascertain that the observed differences are too large to attribute them to random influences only. The MIR community is often criticized for the lack of statistical evaluation it performs, e.g., only two papers in the ISMIR

2004 proceedings (Buyoli and Loureiro 2004) employ a statistical test to measure the statistical significance of their results. A first evaluation benchmark took place at the 2004 ISMIR conference (Cano et al. 2006, unpublished) and ongoing discussions about evaluation of MIR experiments have led to the establishment of the annual evaluation campaign for MIR algorithms ("Music Information Retrieval Evaluation eXchange", MIREX) (Downie 2006). Starting with the MIREX 2006 evaluation (Downie 2006), statistical tests have been used to analyze results in most tasks. But besides using the proper instruments to establish the statistical significance of results, it is equally important to make sure to control all important factors in the experimental design, always bearing in mind that statistical significance does not measure practical importance for users (Johnson 1999; Urbano et al. 2012).

In 2012, MIREX consisted of 15 tasks, such as audio classification, melody extraction, audio key detection to structural segmentation and audio tempo estimation. All these tasks follow a systems-based evaluation framework, in which we mainly measure different characteristics of the system response. The only user component included in these evaluations is the ground truth data, which usually consists of very low-level annotations such as beat marks, tempo, frequency, etc. The two exceptions that include a high-level form of ground truth, closer to a real-world setting, are *Audio Music Similarity and Retrieval* and *Symbolic Melodic Similarity*, in which human listeners provide annotations regarding the musical similarity between two music clips. But it is very important to realize that the real utility of a system for a real user goes far beyond these simple expected-output annotations and effectiveness measures, no matter how sophisticated they are (Ingwersen and Järvelin 2005; Mizzaro 1997). Systems-based evaluations, as of today, completely ignore user context and user properties, even though they clearly influence the result. For example, human assessors in the similarity tasks provide an annotation based on *their* personal and subjective notion of similarity. Do all users agree with that personal notion? Definitely not, and yet, we completely ignore this fact in our systems-based evaluations.

The situation concerning evaluation of user-centric MIR research is far less well developed. In a recent comprehensive review (Weigl and Guastavino 2011) of user studies in the MIR literature by Weigl and Guastavino, papers from the first decade of ISMIR conferences and related MIR publications were analyzed. A central result is that MIR research has a mostly systems-centric focus. Only twenty papers fell under the broad category of "user studies", which is an alarmingly small number given that 719 articles have been published in the ISMIR conference series alone. To make things worse, these user studies are "predominantly qualitative in nature" and of "largely exploratory nature" (Weigl and Guastavino 2011). The explored topics range from user requirements and information needs, insights into social and demographic factors to user-generated meta-information and ground truth. This all points to the conclusion that evaluation of user-centric MIR is at its beginning and that especially a more rigorous quantitative approach is still missing.

3.2 A closer look at the music similarity tasks

In discussing the challenges of quantitative evaluation of user-centric MIR we like to turn to an illustrative example: the recent 2012 *Audio Music Similarity and Retrieval* (AMS) and *Symbolic Melodic Similarity* (SMS) tasks[3] within the annual

[3]The MIREX 2012 results and details can be found at http://www.music-ir.org/mirex/wiki/2012.

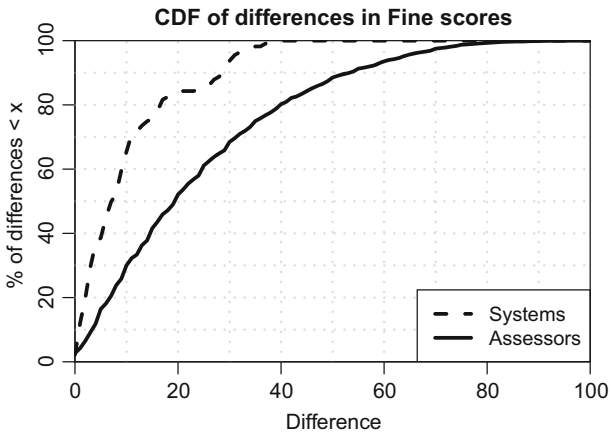

Fig. 2 Distribution of differences among MIREX 2006 AMS assessors and among participating systems in 2006, 2007, 2009, 2010, 2011 and 2012. Differences among assessors are larger than differences among systems

MIREX (Downie 2006) evaluation campaign. In the AMS task, each of the competing algorithms was given 50 random queries (five from each of ten different genres), while in the SMS task each system was given 30 queries. All systems had to rank the songs in a collection (7,000 30-second-audio-clips in AMS and 5,274 melodies in SMS) according to their similarity to each of the query songs. The top ten songs ranked for each query were then evaluated by human graders. For each individual (query, candidate)-pair, a single human grader provided both a Fine score (from 0 to 100) and a Broad score (not similar, somewhat similar, or very similar) indicating how similar the songs were in *their* opinion. The objective here is again to compare all systems (the treatments); the dependent variable is the aggregated score of the subjects' Broad and Fine appraisal of the perceived similarity. From these scores over a sample of queries, we estimate the expected effectiveness of each system for an arbitrary query, and determine which systems are better accordingly.

But since this is a real-world experiment involving human subjects, there is a whole range of additional factors that influence the results. For instance, there are social and demographic factors that might clearly influence the user's judgment of music similarity: their age, gender, cultural background, and especially their musical history, experience, and knowledge. But also factors concerning their momentary situation during the actual listening experiment might have an influence: time of day, mood, physical condition, etc. Not to forget more straightforward variables like type of speakers or headphones used for the test. It is clear that all these variables influence the perceived similarity between two songs and thus the system comparisons, but none of them is considered in the experiments.

In the 2006 run of MIREX, three different assessors provided similarity annotations for the AMS and SMS tasks (Jones et al. 2007). As expected, there were wide differences between assessors, most probably due to their different context and background characteristics. As Fig. 2 shows, over 50 % of the times there was a difference larger than 20 between the Fine scores given by two of the AMS assessors, and even large differences over 50 were observed more than 10 % of the times. This indicates that differences between end users can be quite large, which is particularly worrying considering that observed differences among systems are much smaller (e.g., the difference between the best and worst 2012 systems was just 17, again according to the Fine scale). In fact, recent work established that as much as 20 %

of the users are not satisfied by system outputs which were supposed to be "perfect" according to the systems-based evaluations (Urbano et al. 2012). That is, as much as 20 % improvement could be achieved if we included the user context and user properties as part of our queries so that systems personalize their outputs. But what we actually do in these experiments is ignore these user effects, so we should at best consider our human assessors as a sample from a wider population.[4] As such, we can only interpret our results as the expected performance of the systems, not only for an arbitrary query, but also for an arbitrary user. If we want to evaluate our systems in a more realistic setting, we must change the queries from "what songs are similar to this one" to "what songs are similar to this one, *if we target a user like this*".

As mentioned in Section 1, even the choice of dependent variable is debatable. After all, what does "similar" really mean in the context of music? Timbre, mood, harmony, melody, tempo, etc. might all be valid criteria for different people to assess similarity. This points to a certain lack of rigor concerning the instruction of subjects during the experiment. Also, is similarity the only variable we should measure? The system–user interaction can be characterized with many more variables, some of which are not related to similarity at all (e.g., system response time, ease of use or interface design) (Hu and Kando 2012). Furthermore, the relationship between a system-measure and a user-measure might not be as we expect. For instance, it has been shown that relatively small differences in systems-based measures such as similarity are not even noticed by end users, questioning the immediate practical significance of small improvements and showing the need for systems-based measures that more closely capture the user response (Urbano et al. 2012).

This enumeration of potential problems is not intended to badmouth these MIREX tasks, which still are a valuable contribution and an applaudable exception to the rule of low-level, nearly algorithm-only evaluation. But it is meant as a warning, to highlight the explosion of variables and factors that might add to the variance of observed results and might obscure significant differences. In principle, all such factors have to be recorded at the least, and provided to the systems for better user-aware evaluations. If MIR is to succeed in maturing from purely systems-based to user-centric research, we will have to leave the nice and clean world of our computers and face the often bewilderingly complex real world of real human users and all the challenges this entails for proper design and evaluation of experiments. To make this happen it will be necessary that our community, with a predominantly engineering background, opens up to the so-called "soft sciences" of psychology and sociology, for instance, which have developed instruments and methods to deal with the complexity of human subjects.

4 What should we do?

Incorporating real users in both the development and assessment of music retrieval systems is of course an expensive and arduous task. However, recent trends in music distribution, in particular the emergence of music streaming services that make available millions of tracks to their users, call for intelligent, personalized and context-aware systems to deal with this abundance. Concerning the development

[4]Even though this is likely not the case in the MIREX AMS and SMS tasks as the judgments are certainly biased towards that of music researchers and scientists.

of such systems, we believe that the following three reasons have prevented major breakthroughs so far: (i) a general lack of research on user-centric systems, (ii) a lack of awareness of the limitations and usefulness of systems-based evaluation, (iii) the complexity and cost of evaluating user-centric systems. In designing such systems, the user should already be taken into account at an early stage during the development process, and play a larger role in the evaluation process as well. We need to better understand what the user's individual requirements are and address these requirements in our implementations. Otherwise, it is unlikely that even the spiffiest personalized systems will succeed (without frustrating the user). We hence identify the following four key requirements for elaborating user-centric music retrieval systems:

User models that encompass different social scopes are needed. They may aggregate an individual model, an interest group model, a cultural model, and a global model. Furthermore, the user should be modeled as comprehensively as possible, in a fine-grained and multifaceted manner. With today's sensor-packed smartphones, other intelligent devices, and frequent use of social media it has become easy to perform extensive context logging. Of course, privacy issues must also be taken seriously.

Learning more about the real user needs, such as information or entertainment need is vital to elaborate respective user models. To give some examples of aspects that may contribute to these needs, Pu et al. (2011) and Schedl et al. (2012) mention, among others, similarity and diversity, familiarity, novelty, trendiness, attractiveness, serendipity, popularity, enjoyability, and context compatibility.

Personalization aspects have to be taken into account. In this context, it is important to note the highly subjective, cognitive component in the understanding of music and judgment of its personal appeal. Therefore, designing user-aware music applications requires intelligent machine learning and information retrieval techniques, in particular, preference learning approaches that relate the user context to concise and situation-dependent music preferences.

Multifaceted similarity measures that combine different feature categories (music content, music context, user context, and user properties) are required. The corresponding representation models should then not only allow to derive similarity between music via content-related aspects, such as beat strength or instruments playing, or via music context-related properties, such as the geographic origin of the performer or a song's lyrics, but also to describe users and user groups in order to compute listener-based features and similarity scores. Based on these user-centric information, novel personalized and context-aware music recommender systems, retrieval algorithms, and music browsing interfaces will emerge.

Evaluation of user-centric music retrieval approaches should include in the experimental design all independent variables that are able to influence the dependent variables. In particular, such factors may well relate to individual properties of the human assessors, which may present problems of both practical and computational nature.

Furthermore, it is advisable to make use of recent approaches to minimize the amount of labor required by the human assessors, while at the same time maintaining

the reliability of the experiments. This can be achieved, for instance, by employing the concept of "Minimal Test Collections" (MTC) (Carterette et al. 2006) in the evaluation of music retrieval systems (Urbano and Schedl 2013). The idea of MTC is that there is no need to let users judge all items retrieved for a particular query in order to estimate with high confidence which of two systems performs better. Analyzing which queries (and retrieval results) are the most discriminative in terms of revealing performance differences between two systems, it is shown in Urbano and Schedl (2013) that the number of user judgments can be reduced considerably for evaluating music retrieval tasks.

When looking at user-centric evaluation in fields related to MIR, it seems that in particular the Text-IR and Recommendation Systems communities, are already a step further. They especially foster the use of evaluation strategies that result in highly specific qualitative feedback on user satisfaction and similar subjective demands, for instance in Pu et al. (2011), and Dooms et al. (2011). Such factors are unfortunately all too frequently forgotten in MIR research. We should hence broaden our view by looking into how other communities address the user, investigate which strategies can also be applied to our tasks, and what we can thus borrow from these communities. For example, user aspects reported in Pu et al. (2011), Dooms et al. (2011), and Schedl et al. (2012) include perceived similarity, diversity, familiarity, novelty, trendiness, attractiveness, serendipity, popularity, enjoyability, transparency, and usefulness. We presume that at least some of these also play an important role in music retrieval and should thus be considered in user-centered evaluation of MIR systems.

By paying attention to these advices, we are sure that the exciting field of user-centric music information retrieval will continue to grow and eventually provide us with algorithms and systems that offer personalized and context-aware access to music in an unintrusive way.

References

Aucouturier, J.J., & Bigand, E. (2012) Mel Cepstrum & Ann Ova: The difficult dialog between MIR and music cognition. In *Proc. ISMIR*.

Aucouturier, J.J., & Pachet, F. (2003). Representing musical genre: a state of the art. *Journal of New Music Research, 32*(1), 83–93.

Azzopardi, L., Järvelin, K., Kamps, J., Smucker, M.D. (2011). Report on the SIGIR 2010 workshop on the simulation of interaction. *SIGIR Forum, 44*(2), 35–47.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern information retrieval – the concepts and technology behind search* 2nd edn. Addison-Wesley, Pearson, Harlow, England.

Baltrunas, L., Kaminskas, M., Ludwig, B., Moling, O., Ricci, F., Lüke, K.H., Schwaiger, R. (2011). InCarMusic: context-aware music recommendations in a car. In *Proc. EC-Web*.

Biehl, J.T., Adamczyk, P.D., Bailey, B.P. (2006). DJogger: A mobile dynamic music device. In *CHI 2006: Extended Abstracts*.

Bogdanov, D., & Herrera, P. (2011). How much metadata do we need in music recommendation? A subjective evaluation using preference sets. In *Proc. ISMIR*.

Buyoli, C., & Loureiro, R. (2004). Fifth international conference on music information retrieval. Universitat Pompeu Fabra. http://books.google.at/books?id=r0BXAAAACAAJ.

Callan, J., Allan, J., Clarke, C.L.A., Dumais, S., Evans, D.A., Sanderson, M., Zhai, C. (2007). Meeting of the MINDS: an information retrieval research agenda. *SIGIR Forum, 41*(2), 25–34.
Carterette, B., Allan, J., Sitaraman, R. (2006). Minimal test collections for retrieval evaluation. In *Proc. SIGIR* (pp. 268–275). Seattle, WA, USA.
Cebrián, T., Planagumà, M., Villegas, P., Amatriain, X. (2010). Music recommendations with temporal context awareness. In *Proc. RecSys*.
Clarke, C., Kolla, M., Cormack, G., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I. (2008). Novelty and diversity in information retrieval evaluation. In *Proc. SIGIR* (pp. 659–666). Singapore.
Cunningham, S., Caulder, S., Grout, V. (2008). Saturday night or fever? context-aware music playlists. In *Proc. Audio Mostly*.
Cunningham, S.J., Bainbridge, D., Falconer, A. (2006). More of an art than a science: Supporting the creation of playlists and mixes. In *Proc. ISMIR* (pp. 474–477). Victoria, Canada.
Cunningham, S.J., Downie, J.S., Bainbridge, D. (2005). "The pain, the pain": Modelling music information behavior and the songs we hate. In *Proc. ISMIR* (pp. 474–477). London, UK.
Cunningham, S.J., Jones, M., Jones, S. (2004). Organizing digital music for use: An examination of personal music collections. In *Proc. ISMIR* (pp. 447–454). Barcelona, Spain.
Dooms, S., De Pessemier, T., Martens, L. (2011). A user-centric evaluation of recommender algorithms for an event recommendation system. In *Proc. RecSys: Workshop on Human Decision Making in Recommender Systems (Decisions@RecSys'11) and User-Centric Evaluation of Recommender Systems and Their Interfaces (UCERSTI 2)* (pp. 67–73). Chicago, IL, USA.
Dornbush, S., English, J., Oates, T., Segall, Z., Joshi, A. (2007). XPod: A human activity aware learning mobile music player. In *Proc. Workshop on Ambient Intelligence, IJCAI*.
Downie, J.S. (2006). The Music Information Retrieval Evaluation eXchange (MIREX). D-Lib Magazine. http://dlib.org/dlib/december06/downie/12downie.html.
Elliott, G.T., & Tomlinson, B. (2006). PersonalSoundtrack: context-aware playlists that adapt to user pace. In *CHI 2006: Extended Abstracts*.
Firan, C.S., Nejdl, W., Paiu, R. (2007). The benefit of using tag-based profiles. In *Proceedings of the 5th Latin American Web Congress (LA-WEB)* (pp. 32–41). Santiago de Chile, Chile.
Flexer, A. (2006). Statistical evaluation of music information retrieval experiments. *Journal of New Music Research, 35*(2), 113–120.
Futrelle, J. & Downie, J.S. (2003). Interdisciplinary research issues in music information retrieval: ISMIR 2000–2002. *Journal of New Music Research, 32*(2), 121–131.
Hargreaves, D.J., MacDonald, R., Miell, D. (2005). *Musical Communication, Chap. How Do People Communicate Using Music?* Oxford University Press.
Hu, X., & Kando, N. (2012). User-centered Measures vs. system effectiveness in finding similar songs. In *Proc. ISMIR*. Porto, Portugal.
Ingwersen, P., & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.
Järvelin, K. (2012). IR research: systems, interaction, evaluation and theories. *SIGIR Forum, 45*(2), 17–31.
Johnson, D. (1999). The insignificance of statistical significance testing. *Journal of Wildlife Management, 63*(3), 763–772.
Jones, M.C., Downie, J.S., Ehmann, A.F. (2007). Human similarity judgments: Implications for the design of formal evaluations. In *Proc. ISMIR*. Vienna, Austria.
Kamalzadeh, M., Baur, D., Möller, T. (2012). A survey on music listening and management bahaviours. In *Proc. ISMIR* (pp. 373–378). Porto, Portugal.
Kaminskas, M., & Ricci, F. (2011). Location-adapted music recommendation using tags. In *Proc. UMAP*.
Knees, P., Pohle, T., Schedl, M., Widmer, G. (2007). A music search engine built upon audio-based and web-based similarity measures. In *Proc. SIGIR*.
Knees, P., & Widmer, G. (2007). Searching for music using natural language queries and relevance feedback. In *Proc. AMR*.
Knijnenburg, B.P., Schmidt-Thieme, L., Bollen, D.G. (2010). Workshop on user-centric evaluation of recommender systems and their interfaces. In *Proc. RecSys* (pp. 383–384). Barcelona, Spain.
Kohonen, T. (2001). *Self-Organizing Maps, Springer Series in Information Sciences*, vol. 30, 3rd edn. Berlin, Germany: Springer.
Koren, Y., Bell, R., Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer, 42*, 42–49.

Law, E., & von Ahn, L. (2009). Input-agreement: A new mechanism for collecting data using human computation games. In *Proc. CHI* (pp. 1197–1206). Boston, MA, USA.
Lee, D.D., & Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature, 401*(6755), 788–791.
Lee, J.S., & Lee, J.C. (2007). Context awareness by case-based reasoning in a music recommendation system. In *Proc. UCS*.
Liem, C.C., Müller, M., Eck, D., Tzanetakis, G., Hanjalic, A. (2011). The need for music information retrieval with user-centered and multimodal strategies. In *Proc. MIRUM*. Scottsdale, AZ, USA.
Linden, G., Smith, B., York, J. (2003). Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing, 4*(1), 76–80.
Liu, H., Hu, J., Rauterberg, M. (2009). Music playlist recommendation based on user heartbeat and music preference. In *Proc. ICCTD*.
Mizzaro, S. (1997). Relevance: the whole history. *Journal of the American Society for Information Science, 48*(9), 810–832.
Moens, B., van Noorden, L., Leman, M. (2010). D-jogger: syncing music with walking. In *Proc. SMC*.
Novello, A., McKinney, M.F., Kohlrausch, A. (2006). Perceptual evaluation of music similarity. In *Proc. ISMIR*. Victoria, Canada.
Nürnberger, A., & Detyniecki, M. (2003). Weighted self-organizing maps: incorporating user feedback. In *Proc. ICANN/ICONIP*.
Pohle, T., Knees, P., Schedl, M., Widmer, G. (2007). Building an interactive next-generation artist recommender based on automatically derived high-level concepts. In *Proc. CBMI*.
Pu, P., Chen, L., Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proc. RecSys* (pp. 157–164). Chicago, IL, USA.
Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (Eds.) (2011). *Recommender Systems Handbook*. Springer.
Rocchio, J.J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The SMART Retrieval System - Experiments in Automatic Document Processing* (pp. 313–323). Englewood Cliffs, NJ: Prentice-Hall.
Sarwar, B., Karypis, G., Konstan, J., Reidl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proc. WWW*.
Schedl, M., Hauger, D., Schnitzer, D. (2012). A model for serendipitous music retrieval. In *Proc. IUI: 2nd International Workshop on Context-awareness in Retrieval and Recommendation (CaRR 2012)*. Lisbon, Portugal.
Schedl, M., & Knees, P. (2011) Personalization in multimodal music retrieval. In *Proc. AMR*.
Tunkelang, D. (2008). Transparency in Information Retrieval. http://thenoisychannel.com/2008/08/27/transparency-in-information-retrieval. Accessed Aug 2008.
Urbano, J., Downie, J.S., McFee, B., Schedl, M. (2012). How significant is statistically significant? The case of audio music similarity and retrieval. In *Proc. ISMIR* (pp. 181–186). Porto, Portugal.
Urbano, J., & Schedl, M. (2013). Minimal test collections for low-cost evaluation of audio music similarity and retrieval systems. *International Journal of Multimedia Information Retrieval, 2*(1), 59–70.
Weigl, D., & Guastavino, C. (2011). User studies in the music information retrieval literature. In *Proc. ISMIR*.
Woelfer, J.P., & Lee, J.H. (2012). The role of music in the lives of homeless young people: A preliminary report. In *Proc. ISMIR* (pp. 367–372). Porto, Portugal.
Wolff, D., & Weyde, T. (2011). Adapting metrics for music similarity using comparative ratings. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)* (pp. 73–78). Miami, FL, USA.
Xu, W., Liu, X., Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proc. SIGIR*.
Xue, G.R., Han, J., Yu, Y., Yang, Q. (2009). User language model for collaborative personalized search. *ACM Transactions on Information Systems, 27*(2), 11:1–11:28.
Zhang, B., Shen, J., Xiang, Q., Wang, Y. (2009). CompositeMap: A novel framework for music similarity measure. In *Proc. SIGIR*.
Zhang, Y., Callan, J., Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In *Proc. SIGIR* (pp. 81–88). Tampere, Finland.