People-Powered Music: Using User-Generated Tags and Structure in Recommendations

Johannes Schleith
Institute of Cognitive Science
University of Osnabrück
49069 Osnabrück, Germany
jschleit@uni-osnabrueck.de

Simone Stumpf
Centre for HCI Design
City University London
London EC1V 0HB, UK
Simone.Stumpf.1@city.ac.uk

Todd Kulesza
School of EECS
Oregon State University
Corvallis, Oregon 97333, USA
kuleszto@eecs.oregonstate.edu

ABSTRACT
Music recommenders often rely on experts to classify song facets like genre and mood, but user-generated folksonomies hold some advantages over expert classifications—folksonomies can reflect the same real-world vocabularies and categorizations that end users employ. We present an approach for using crowd-sourced common sense knowledge to structure user-generated music tags into a folksonomy, and describe how to use this approach to make music recommendations. We then empirically evaluate our "people-powered" structured content recommender against a more traditional recommender. Our results show that participants slightly preferred the unstructured recommender, rating more of its recommendations as "perfect" than they did for our approach. An exploration of the reasons behind participants' ratings revealed that users behaved differently when tagging songs than when evaluating recommendations, and we discuss the implications of our results for future tagging and recommendation approaches.

Categories and Subject Descriptors
H.1.2 [**Models and Principles**]: User/Machine Systems - *Human information processing;* H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval - *Information filtering*

General Terms
Algorithms; Human Factors.

Keywords
Music recommender systems; folksonomies; user-generated tags.

1. INTRODUCTION
Systems that personalize themselves to end users are becoming commonplace, particularly music recommenders—Pandora.com and Last.fm, for example, now have more than 100 million combined listeners enjoying personalized "radio stations" and playlists. While collaborative filtering [9, 24] can power recommenders by grouping similar *users* together to personalize suggestions, approaches that suggest songs based on the *content* of the music are also being explored [22]. Increasingly, expert-led descriptions of songs, such as Pandora's Music Genome project [19], and automatically generated tags [1] are being replaced by integrating user-generated data [7, 17] into recommendations. One major advantage of user-generated tags is that they reflect the vocabulary end users employ to describe certain fields, such as songs and artists on Last.fm. This dynamic character enables them to quickly adapt to changes and newly evolving trends. Further, user-generated tags are based on the "wisdom of the crowds": instead of being curated by a single authority or group of specialists, every user has the opportunity to contribute to this collective knowledge base.

Relying on user-generated tags, however, is not without issues. Tags are loose, collaboratively created collections of words, and hence may contain ambiguous, synonymous, or idiosyncratic descriptions. Further, they can have different meanings under different listening contexts or for different users. Since these user-generated tags also lack structure, using them in recommendations becomes especially difficult. Structuring collections of user-generated tags aims to turn these collections into folksonomies (i.e., user-generated ontologies), thus increasing their potential utility for retrieval and recommendation systems.

This paper describes a new approach (illustrated in Figure 1) for combining user-generated song tags with crowd-sourced common-sense concepts to produce *concept-clustered recommendations*. We evaluated our approach via a user study, comparing concept-clustered recommendations with an unstructured content-based recommendation approach. We also elicited participant explanations of their own criteria for determining appropriate and inappropriate recommendations.

Formally, this work investigates the following four research questions:

RQ 1. What are the contents of a music-related folksonomy (i.e., what do people talk about when they are tagging)?

RQ 2. Can a crowd-sourced ontology accurately structure user-generated tags, and if so, how can the resulting structure be integrated into recommendations?

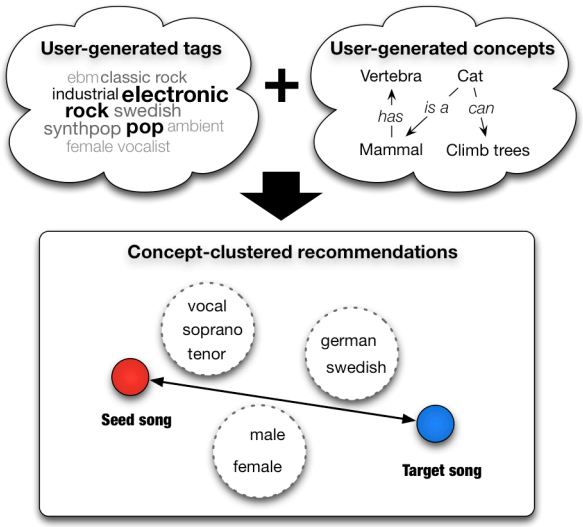

Figure 1: We structured music tags via a user-generated ontology and used the resulting clusters to make music recommendations.

RQ 3. How do music recommendations based on structured tags compare to recommendations based on unstructured user tags?

RQ 4. What do end users want to tell music recommenders about song similarity? How well does this rich user feedback align with existing user-generated tags?

Our results hold lessons for structuring user-generated music tags, for incorporating structural properties of tags into music recommenders, and for the design of systems and interfaces that support tagging and recommendations.

2. RELATED WORK

2.1 How People Describe Music

The language that end users employ and the facets they discuss when dealing with music have only recently begun to be investigated. Studies conducted with DJs have indicated the need for more expressive descriptions of songs than are currently available through meta-tags [3]. Similarly, professionals who look for music to add to films, TV commercials, and computer games employ common musical facets like artist, time period, and tempo, and also emotive aspects such as mood [10]. Studies of musically untrained users have found that artist, genre, event/activity, mood, and tempo are prominent facets when putting together a playlist [4, 8, 29]. Recent research has also tried to understand user-generated tags for music in terms of the distinct facets that they describe [2].

Current approaches investigating how people describe music rely on facet classifications by experts, manually applying these facets to user-generated tags. Because folksonomies constantly evolve, such manual approaches for structuring user-generated tags quickly become infeasible for many real-world applications; the structuring process needs to performed automatically if it is to keep pace with a changing folksonomy.

2.2 Automatically Structuring Song Tags

Some prior work in music systems has attempted to automatically structure user-generated tags, but none of the existing approaches are yet suitable for the wide variety of facets discussed in such tags. One solution, for example, is to reduce a tag space's dimensionality via latent semantic analysis (LSA) [17]. The dimensions derived from LSA, however, tend to heavily focus on only two musical facets: artist and genre [17].

Other approaches have used existing ontologies to automatically map user-generated tags to related musical facets, in part inspired by encouraging results from using ontologies (e.g., WordNet) to structure other types of multimedia data [27]. The work in this area has focused on using expert-curated ontologies (e.g., Wikipedia) to provide the facets and hierarchy for grouping user-generated tags [25, 34]. Because these ontologies still rely on expert users to provide the general classification system, crowd-sourcing has no impact on their overall structure. In this paper, conversely, we investigate the viability of automating the structuring step, leveraging a fully crowd-sourced ontology to reveal the folksonomy underlying a collection of otherwise unstructured tags.

2.3 Telling a Recommender How to Behave

Many systems provide rich ways for users to describe songs via tags, but support only limited mechanisms for users to provide feedback about recommendations. Richer forms of feedback to recommenders and machine learning systems have received recent interest, including ways for end users to understand and control predictions [14, 30], direct modification of a classifier's cost matrix [11], and user-directed creation of novel ensemble classifiers [32]. Researchers have also studied the human costs (e.g., time to complete, cognitive load) of various feedback mechanisms [26], how an understanding of a recommender's reasoning can impact users' perceived cost/benefit tradeoff of providing such feedback [13], and the appropriateness of different interaction techniques, depending on a user's goals or personal characteristics [12]. This paper explores whether tagging capabilities, when leveraged by recommender systems, can provide a source of rich feedback to better personalize recommendations.

3. OUR APPROACH

The core of our approach is the application of a crowd-sourced collection of real-world knowledge to structure user-generated descriptions of music. This section explains how we harvested tags, the musical facets these tags discuss, how we automatically structured the tags using ConceptNet, and how we used the resulting structured information to power a music recommender.

3.1 User-Generated Music Tags

First, we needed to collect a sample of user-generated music tags. We used the Audioscrobbler 2.0 API [15] to access tags listeners applied to Last.fm songs, but because Last.fm does not support direct access of their entire tag space, we needed to proceed in stages. We first retrieved the 250 most-frequently assigned tags for all songs, and then collected the top 250 songs associated with each of these tags. From each song, we extracted every tag that users had applied (as many as 30 tags per song). The resulting corpus consisted of 132,118 user-generated tags for 51,618 distinct songs.

To answer RQ1, we randomly extracted a sample of 500 tags and categorized them by the musical facets each tag described. We based our classification scheme on prior research [21, 29], refining it via the grounded theory approach (i.e., some categories were combined, expanded, or created based on the data we encountered). In order to assess the reliability of the resulting facets, two researchers independently coded 300 tags (out of the sample of 500). Their inter-rater reliability, as measured by Cohen's Kappa, was 0.61 over the entire collection of facets, indicating satisfactory agreement. The 12 facets we identified are listed in Table 1.

Table 1 also shows how often each facet was encountered in our tag sample. Our analysis revealed that almost one quarter (23%) of the tags were too ambiguous to be classified at all (tags such as *One tag to rule them all*). Another 22% directly referred to the song's artist, title, album, or composer, without revealing any additional information that could not be extracted from an audio file's metadata. Assuming that a recommender's source audio files include complete metadata, the implication is that an automatic analysis of user-generated tags may not contribute additional useful information for nearly half of the available tags.

The remaining tags, however, largely discussed facets that would be otherwise unavailable to a recommender. Our analysis showed that end users frequently used tags to express their emotional response to a song (Mood: 12%) and enthusiasm for it (Rating: 10%), as well as their determination of a track's Genre (13%). Genre tags reflect a *user's* determination of category, which may differ from expert-assigned metadata; particularly among subcultures, there may be additional value in precise genre tags

Facet	Definition	Examples	Count	% of tags
Artist	Naming or referring to a composer, piece or album.	*Beatles cover*	112	22.4%
Genre	Naming a genre or style of music.	*Rock, Ska*	69	13.8%
Mood	Expressing the listener's feelings or the mood of the listener.	*Lonely, angry, I have to laugh, gives me power*	60	12.0%
Rating	Enthusiasm about or rating of a song.	*Best song, awesome, I like*	51	10.2%
Instrument/Acoustics	Description of instrumentation or performance characteristics.	*Female voice, guitar solo*	27	5.4%
Environment	Appropriate locations or environments for a song.	*Travelling, while driving, sunset, Friday night*	25	5.0%
Subjective	Referring to the user rather than the music.	*Albums I own, DVD I'd like to have*	19	3.8%
Sexual Allusions	Sexual allusions and wordplay.	*Sexy, I have to change my underwear after this song*	6	1.2%
Location	Referencing the geographic or political location where the music originates.	*African, South American, Germany, Hamburg*	6	1.2%
Era	Temporal information about a year, period, or style.	*00s, middle ages, baroque*	4	0.8%
Culture	Information about an ethnic group or culture.	*Celtic, Christian*	3	0.6%
Tempo/Rhythm	Referring to the tempo or rhythm of a song.	*Strong beat, driving rhythm*	0	0.0%
Other	Tags that do not fit into any of the above categories.	*One tag to rule them all*	118	23.6%

Table 1. Facets for a random sample of 500 user-generated tags, ordered in decreasing frequency.

that reflect a culture's current vocabulary and stratification. Other facets mentioned included a song's instrumentation (Instrumentation/Acoustics: 5%) and ideal listening environments (Environment: 5%).

Not all problems are appropriate domains for crowd-sourced solutions, and we identified three musical facets where user-generated tags were of little value. We found that listeners rarely use tags to describe geographical locations (Location: 1%) or the musical era of a song (Era: <1%), and surprisingly, not even one tag in our sample described a song's tempo or rhythm (Tempo/Rhythm: 0%). However, as facets like Era and Location can be extracted from other sources (such as an album's release date or a band's hometown), and Tempo/Rhythm can often be determined from an audio file's waveform, we believe the results of this analysis can help inform system developers which machine-learning features to elicit via the "wisdom of the crowd", versus those that need to be gathered using complementary techniques.

Additional difficulties for an automatic, natural language approach to classification arise due to accidental misspelling (e.g., the tag *Sill Dre* was applied to the song *Still Dre*), intentional misspellings based on cultural constraints (e.g., the use of "leet"-speak for electronic music tags like *DR4MNB4SS*), and neologisms (e.g., *Favoritized*). Further, because Last.fm is an international service, their tags encompass the languages of a diverse user base (e.g., *Schweineorgel* is colloquial for "accordion" in German, *Bateria* is Spanish for "drum kit").

Our analysis of the Last.fm tags also revealed a surprising pattern—many tags were employed as personal annotations rather than descriptions of a song's content. Tags such as *I love Canada* and *There for you* are illustrative examples of the personal statements we observed in our sample. Recommendation approaches seeking to leverage user-generated tags will need to account for the manner in which end users repurpose tags, such as discarding those that provide no discriminatory value, or transforming them into a more consistent (and hence, comparable) format.

3.2 Structuring Tags via ConceptNet

Our second research question (RQ2) explores the difficulties of combining two crowd-sourced knowledge bases, particularly the problem of structuring user-generated tags via a lightweight, user-generated semantic net. Structuring via semantic networks is not as straightforward as structuring via an ontology like Wikipedia, as semantic networks do not have a strict hierarchy of categories. We next present an algorithm for structuring tags using ConceptNet and AnalogySpace (integrating the resulting structure into recommendations will be discussed in Section 3.3).

ConceptNet [18] is a crowd-sourced, common-sense semantic network; it extracts concepts and relationships (Figure 2) from sentences people type into the Open Mind Common Sense Project [20]. To interact with ConceptNet, we used AnalogySpace [28] and the Divisi [5] package, which provide an efficient representation of the network and tools to manipulate it.

ConceptNet and AnalogySpace provide measures about the *similarity* and *relatedness* of two concepts. Similarity is based on the "closeness" of two concepts in a vector space, as determined by their shared features. Relatedness, in contrast, uses spreading activation, which models the "spread" of energy from one concept to its neighbors via connecting relationships. The combination of similarity and relatedness allowed us to model the complex connections between concepts; for example, "sad" and "cry" are closely *related*, even though they are not very *similar*.

As discussed in Section 3.1, user-generated tags require preprocessing to "clean" them prior to structuring. Our cleaning processes involved the following steps:

1. Remove artist and song titles from the tag collection
2. Remove stopwords (e.g., "the", "a", etc.) and punctuation
3. Transform compound tags into single words
4. Correct misspellings by condensing repeated characters (e.g., "cuuuuuuute" would become "cute")
5. Correct remaining misspellings with the word the smallest Levenshtein distance away (up to a max of 2.0)
6. Translate "leetspeak" into English (e.g., "v01c3" would become "voice")
7. Remove any tags which do not have a corresponding concept node in ConceptNet

The result of this process was a set of input tags for ConceptNet.

We exploited ConceptNet's inherent organization—via concepts and semantic relationships—to structure tags using a *k*-means++ clustering approach. There are two main steps to this algorithm: an *assignment* step in which each tag is assigned to a cluster, and an *update* step in which a new centroid (or medoid) for the cluster is calculated. Instead of starting with random medoids, our variation searches the entire collection for the *k* tags (concepts) with the largest number of relationships to other tags; it then selects these tags as medoids (i.e., the initial categories). We repeat the assignment and update steps until the clusters are stable (i.e., when the medoids cease changing by a significant degree with each update). Our implementation experimented with multiple values of *k* (50, 500, and 1,000), finding the best results with $k = 500$.

We experimentally evaluated two different metrics for assigning tags to clusters, individually testing Divisi's similarity and spreading activation measures. We also evaluated three different metrics for updating the medoids: Divisi's similarity and spreading activation measures, plus a novel function we implemented that searches out higher-level classes for the tags in a given cluster. Our new function works by identifying the concepts shared between all tags in a cluster, and then restricts the potential medoid concepts to those on the right-hand side of *is-a* relationships. Figure 3 illustrates an example. We tested each of these measures on a small number of non-musical tags and used the Rand Index [23] to compare the resulting clusters to manual groupings by a researcher. Our best results were achieved by employing spreading activation to assign tags to clusters and our novel function to update the medoids.

We evaluated our automatic structuring approach to quantify how well it could structure a collection of tags versus manual categorization. We compared our algorithm's clusters of the same 500 tags previously classified (see Section 3.1), using the facets from Table 1 as the "true" groupings. Agreement between the manual grouping and our algorithm's clustering yielded a Rand Index of 83.2%, based on a cluster size of two; this can be considered a very satisfactory degree of agreement, as two human coders only achieved a Rand Index of 86.0% on the same dataset and cluster size.

A detailed exploration of the automatic structuring, however, revealed that ConceptNet might encounter some problems unique to the music domain. When our approach differed from the structuring by human coders, it was usually the specific domain context that mattered. For example, the tag "rock" was understood

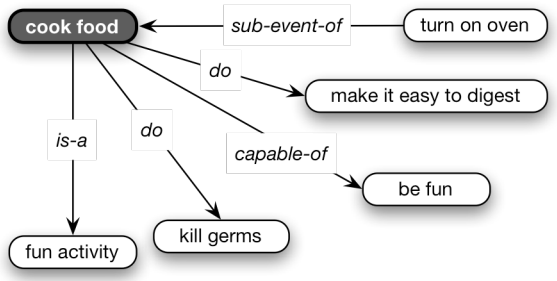

Figure 2. ConceptNet provides a repository of concepts and their relationships to one another.

to be a "cliff" by ConceptNet, but meant a music genre to the human coders (and, presumably, the taggers on Last.fm). Another difficulty resulted from ConceptNet's relative data sparseness regarding musical terms. For example, "Ska" (a genre that has existed for decades) was not represented within the sematic network. Furthermore, our algorithm often adopted overly broad top-level categories. For example, while a human may separate two tags into Mood and Instrument/Acoustics, the algorithm would sometimes cluster both together under the broad concept "Music". Each of these issues may result from the poor coverage of music-related concepts currently available in ConceptNet; end users have, thus far, provided fewer details about music than other "common sense" areas.

3.3 Concept-Clustered Recommendations

Once the work of cleaning and structuring user-generated tags is complete, the result can be used to power a content-based recommender system. We first built a tag matrix *T*, whose rows were the 51,618 songs we harvested from Last.fm, and whose columns were the harvested tags themselves (see Section 3.1). The cells of matrix *T* held a value of "1" if the corresponding tag was applied to the given song, or a "0" otherwise. We also built a structure matrix *S*, whose rows and columns were the tags; for each pair of tags that were assigned to the same cluster (based on the structuring described in Section 3.2), matrix *S* held their similarity value, while pairs of tags that were *not* clustered together held values of "0".

The matrices *S* and *T* allowed us to build a recommender system that "knows" about tag structure. For any two songs, our concept-clustered recommender first retrieves the list of tags *c* common to both songs from matrix *T*. It then calculates the distance *d* between the two songs by summing the values in *S* for the tags in *c*. The distance *d* is normalized by the number of tags in *c* and the total number of tags applied to the two songs. To make

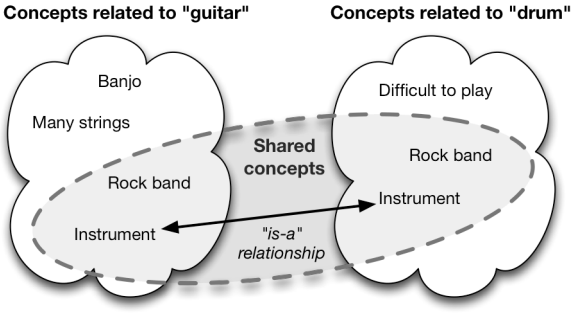

Figure 3. We established higher-level classes of tags by identifying "is-a" relationships shared between two groups of concepts.

recommendations, *d* is calculated for every song in the matrices against the seed song; this list is then sorted, and the top *n* songs are returned as recommendations.

4. USER EVALUATION

We evaluated our approach via an empirical study involving real end users, rather than testing it on a "gold standard" data set. This allowed us to directly compare users' ratings of recommendations made by our concept-clustered technique against recommendations made by a content-based system that did not leverage structural information (RQ3), and to investigate participants' explanations of what mattered in recommendations (RQ4).

4.1 Participants and Procedure

Our user study was conducted in the form of an online survey. Potential participants were recruited via announcements on mailing lists, Twitter, Facebook, and LinkedIn, and directed to the survey website; no compensation was offered for their time. Overall, 75 people responded to the survey (49 males and 26 females), ranging in age from 18 to 59. Twenty-six participants completed the entire survey, and 33 completed at least half. Our analysis includes all survey responses from participants, normalized by the number of responses for each question.

On the survey website, respondents began by providing background information (e.g., age, gender, interest in music, etc.). They then listened to a series of 20 pairs of songs. One song in each pair was a seed song, while the other song was the top recommendation for the given seed. After listening to both songs, participants were asked to rate the appropriateness of each recommendation (in relation to its seed song) on a five-point Likert scale and, optionally, describe the reasons for their rating. The tags and concept clusters underlying the recommendations were not visible to participants, as prior work has found that making explanations available carries the risk of influencing the responses to be more positive, and the system to be judged trustworthier, than when explanations are withheld [6].

Our seed songs were the top 20 songs (by unique artists) from the Last.fm Top Charts on 20 July 2011. The user-generated tags for these songs were retrieved as described in Section 3.1, and then cleaned as described in Section 3.2. We conducted our study according to a within-subject design, with each participant exposed to both kinds of recommendations (or conditions), containing a balanced set of ten song pairs. Hence, each seed song was randomly assigned to a condition: in one, it seeded a recommender implementing our concept-clustered approach, while in the other, it seeded a more traditional content-based recommender that computed the similarity of the vectors containing the un-clustered tags, via Divisi. For each recommender, the pool of target songs available as potential recommendations came from a corpus retrieved from Last.fm containing over 51,000 songs (the same collection of songs described in Section 3.1). The full list of seed songs and their associated recommendations is presented in Table 2.

4.2 Were Concept-Clustered Recommendations Better?

Participants rated the appropriateness of each recommendation via a five-point Likert scale. To perform a within-subject comparison, we averaged each participant's ratings for the unstructured recommendations (M = 3.38, SD = 0.79) and concept-clustered recommendations (M = 3.05, SD = 0.87). A paired *t*-test comparing each participant's ratings for unstructured recommendations versus their ratings for concept-clustered recommendations showed that participants consistently rated the unstructured recommendations as more appropriate than concept-clustered recommendations ($t = 2.62$, d.f. = 46, $p = .011$).

Why this discrepancy? Part of the reason involves "perfect" recommendations (i.e., a Likert response of five). While participants rated approximately the same number of recommendations as "awful" (i.e., a Likert response of one) between conditions (48 for the unstructured recommender, versus 51 for the concept-clustered recommender), participants only rated 61 of the concept-clustered recommender's selection as "perfect", while they did so for 102 of the unstructured recommender's choices. Thus, while our concept-clustered recommendations were no more likely to be considered truly awful by participants, they were significantly less likely to be considered perfect (Pearson's Chi-squared test with Yates' continuity correction, $X^2 = 4.43$, d.f. = 1, $p = .035$).

These results suggest that our approach is, at present, slightly worse than using unstructured tags at picking "perfect" songs. We next present an analysis of the reasons underlying these ratings, which provided additional clues as to why these recommendations had trouble reflecting participants' expectations.

4.3 What Makes Recommendations Perfect?

Just as important as *whether* participants found recommendations to be appropriate is *why* they found them to be so. In our study, participants had the option of providing a detailed reason for their rating, giving insight into which aspects of the recommendations mattered to participants. Participants had very diverging ratings; for example, the ratings for even the first recommendation ranged from "perfect" to "awful". The following participant feedback (from that first recommendation) illustrates some of the different reasons informing the range of judgments:

"Both songs got this 80's touch. Similar use of synth."

"Similar melodically and in electronic instrumentation."

"Not even remotely alike. Completely different atmospheres."

"Beats seem to be similar, but the rest isn't at all."

In order to systematically identify the types of things that end users want to tell music recommenders, we analyzed participants' comments to ascertain the musical facets they discussed. Our analysis began with the same list of identified facets from Section 3.1 and employed the same procedure to classify participants' comments. One researcher applied these facets to the comments and, during this process, also identified additional facets that were not covered by the earlier classification scheme. In total, 485 comments were classified. Our final classification scheme is presented in Table 3, alongside the frequency with which participants mentioned each facet in their comments (a single comment could be coded with multiple facets).

Our analysis reveals that the feedback participants gave about the recommendations focused mainly on previously identified facets such as Instrument/Acoustics, Mood, Genre, and Tempo/Rhythm, each of which were mentioned in more than 10% of the comments. The most prevalent new facet was Style (4.2%), which covered comments discussing musical style without explicitly referring to Mood, Tempo/Rhythm, or Genre. Participants also discussed song Popularity (e.g., *"Both pretty popular songs"*, 3.4% of facets mentioned) and intended Audience (e.g., *"I would not recommend the second song to those that like the other song and vice versa. They don't seem to share the same audience."*,

Seed song	Recommended song	Recommender
Lady Gaga – "Judas"	Goldfrapp – "Dreaming"	U
Foster The People – "Pumped Up Kicks"	Broken Bells – "The High Road"	U
Bon Iver – "Perth"	Second Hand Serenade – "Fall for You"	CC
Katy Perry – "Last Friday Night"	Gwen Stefani – "The Sweet Escape"	U
Florence + The Machine – "Dog Days are Over"	Regina Spektor – "On the Radio"	U
Britney Spears – "Hold it Against Me"	Britney Spears – "Till the World Ends"	U
Nirvana – "Smells Like Teen Spirit"	Nirvana – "Come as You Are"	U
Coldplay – "Viva la Vida"	Maroon 5 – "Nothing Lasts Forever"	CC
Jennifer Lopez – "On the Floor"	Britney Spears – "Break the Ice"	U
Rihanna – "S&M"	Hellogoodbye – "Touchdown Turnaround"	U
Mumford & Sons – "The Cave"	MIKA – "Blame It on the Girls"	CC
Arcade Fire – "Ready to Start"	Muse – "Thoughts of a Dying Atheist"	CC
Kings Of Leon – "Sex on Fire"	The Killers – "Mr. Brightside"	CC
MGMT – "Kids"	Modest Mouse – "Dashboard"	U
Oasis – "Wonderwall"	Snow Patrol – "Chasing Cars"	CC
The Last Shadow Puppets – "The Age of the Understatement"	Sum 41 – "Best of Me"	CC
LMFAO – "Party Rock Anthem"	deadmau5 – "FML"	CC
Red Hot Chili Peppers –"Californication"	Red Hot Chili Peppers – "Otherside"	CC
The Strokes – "Under Cover of Darkness"	LCD Soundsystem – "Dance Yrself Clean"	CC
Adele – "Rolling in the Deep"	Amy Winehouse – "Back to Black"	U

Table 2. User study songs and the recommendation approach used, either concept-clustered (CC) or unstructured (U).

1.3% of facets). Music recommenders could potentially exploit this set of facets via richer forms of feedback to help improve their suggestions.

To better understand why our recommendations often failed to please participants, we examined how closely user feedback about the *actual* recommendations matched the user-generated tags underlying *potential* recommendations (recall that these tags, once structured, served as the features our content-based recommender used to determine song similarity). Figure 4 compares how often participants discussed each facet when explaining their ratings with how often those facets were identified in the Last.fm tags (from Section 3.1).

Our results show a stark contrast: what mattered most to participants in recommendations (Instrument/Acoustics and Mood) was entirely different than what people discussed when tagging songs (Artist/Band and Other). In fact, of the top ten facets participants discussed while rating recommendations, only one (Genre) was equally represented in the tag space. Surprisingly, it appears that end users behave very differently when tagging songs than when evaluating recommendations.

The results of our analysis suggest that the data for making "perfect" recommendations does not (yet) exist in user-generated tag spaces. In the next section, we discuss the implications of our findings for the elicitation of musical tags for recommendations.

5. DISCUSSION

Our findings hold implications for researchers working toward improving music recommendations via either the structure of a folksonomy, or the integration of user-generated tags.

5.1 Improving Concept-Clustered Recommendations

Our evaluation of concept-clustered recommendations revealed two primary areas of improvement for our approach: (1) handling tag sparseness, and (2) exploring different clustering and recommendation techniques.

Sparseness is a problem common to both collaborative filtering and content-based recommenders—without sufficient data, recommendation quality will suffer. Our sample, as expected, included a "long tail" of infrequently used tags. One explanation for why participants preferred unstructured recommendations is that by attempting to structure a sparse dataset, we compounded the problem; the structure derived from a limited number of tags may have been too poor to lead to reliably good recommendations. Leveraging additional sources of information (such as identifying a song's tempo via audio analysis) may be necessary to mitigate the negative effects of tag sparseness. Additionally, tag sparseness could be employed to *help* a recommender, as it may reflect a song's popularity (a facet that participants cited 27 times when rating song similarity).

Clearly, structuring techniques play an important role in the resulting recommendations. Our concept-clustered recommendation approach places great importance on reliably structuring tags into "useful" facets, so poor clustering will naturally yield flawed recommendations. Future work should investigate the impact of cluster sizes on the appropriateness of recommendations. Other structuring approaches may also be worth exploring, such as clustering with topic smoothing, which has shown promise in other domains [33].

Additionally, the method for integrating a folksonomy's structure into a recommender matters greatly. Our similarity measure was

Facet	Count	% of facets
Instrument/Acoustics	149	18.9%
Mood	134	17.0%
Genre	105	13.3%
Other	101	12.8%
Tempo/Rhythm	82	10.4%
Artist	51	6.5%
Style	33	4.2%
Popularity	27	3.4%
Rating	26	3.3%
Environment	24	3.0%
Subjective	23	2.9%
Topic	12	1.5%
Audience	10	1.3%
Era	8	1.0%
Location	3	0.4%
Culture	1	0.1%
Sexual Allusions	1	0.1%

Table 3. Classification of the comments participants gave when explaining the recommendations' appropriateness. Newly identified facets are shaded.

based solely on the distance of tags in clusters—future work should explore additional distance or similarity measures. One intriguing alternative to the method presented in this paper (which recommended *n* target songs in decreasing order of similarity) is to use structural information for constructing an entire playlist. Such an approach may be able to mimic informal "rules" of playlist generation (e.g., selecting songs that "flow" together based on facets like tempo or mood [29]) by using tag structure to identify details that tags alone cannot reveal.

5.2 Building a Foundation for People-Powered Music Research

We encountered significant hurdles in the design and implementation of concept-clustered recommendations. Particular challenges were tag retrieval, the use of a lightweight, generic semantic net knowledge base, and a mismatch between user-generated tag content and the information participants wanted recommendations to be based upon.

Although there is currently limited public access to user-generated tags via the Audioscrobbler API, there is no completely public data set from which user-generated tags can be acquired for analysis and testing of recommender systems. Conversely, while ConceptNet *is* a completely public knowledge base, it was precisely its "people-powered" nature that caused problems for us. As mentioned earlier, ConceptNet's "knowledge" of music was far from complete, which negatively impacted our ability to structure the tag space. Finding ways to motivate contributions to public knowledge bases (such as gamification systems like Tagatune [16]) could help alleviate this problem. For now, there remains a strong need for publicly available, crowd-sourced data sets and ontologies in the musical domain.

The final issue identified by our follow-up analysis was the discrepancy between participants' reasons for finding a

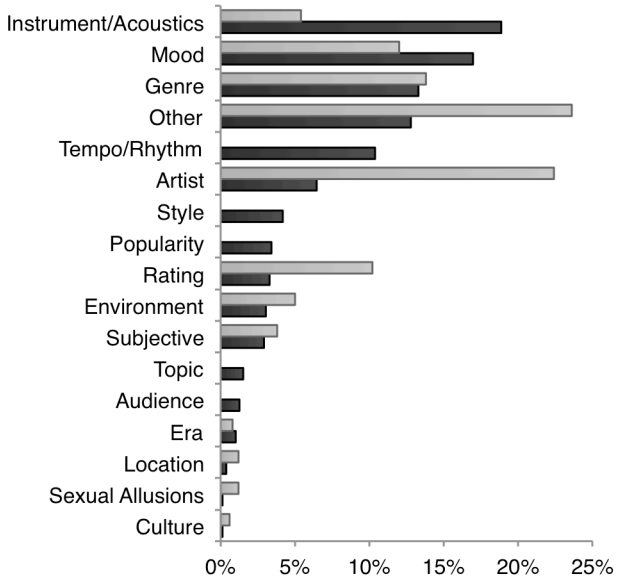

Figure 4. Percentage of musical facets discussed by participants when evaluating recommendations (dark) versus the same facets' prevalence in a random sample of Last.fm tags (light).

recommendation appropriate and the contents of Last.fm tags: when Last.fm listeners tagged songs, they rarely discussed the facets that our participants cared about when evaluating music. This presents a problem for using existing tags in recommendations. Adjusting the tagging interface, however, may resolve this issue for future recommender systems. Priming, for example, has been shown to significantly impact the thoughtfulness of user-contributed comments online [31]. A related approach may be able to subtly steer the content of tags toward topics that participants relied upon when judging music similarity, such as instrumentation and emotional responses. Successfully eliciting such tags will overcome a major hurdle to employing user-generated tags in recommender systems.

6. CONCLUSION

This paper presented an analysis of Last.fm tags, described an approach to leverage ConceptNet to structure these tags, and a method to integrate their structure into recommendations. We evaluated our approach through an empirical study involving real users. Our findings show that:

- **End users repurpose tags in surprising ways:** Although some users do tag song facets such as Mood, Rating, or Genre, tags are often used as personal annotations; nearly half of our sample held no useful information for recommender systems.
- **Crowd-sourced tags can be reliably structured via crowd-sourced concepts:** Combining these two knowledge bases showed promising results, with an automated approach approximating the accuracy of manual structuring techniques.
- **Concept-clustered recommendations are viable, but imperfect:** End user evaluations of concept-clustered recommendations were slightly above average, but still worse than recommendations based on unstructured tags.
- **People attend to aspects of songs differently when tagging them than when listening to recommendations:** For

people-powered music recommenders to excel, tagging interfaces may need to steer users toward providing data that is more useful for recommendations than the annotations they current employ.

Our work points a way forward for "people-powered" recommendations: systems based on both crowd-sourced tags and structure. Our results suggest that there are still significant hurdles to this approach in the music domain, but an increasingly rich space of crowd-sourced musical tags and facets could help to surmount them. Music, of course, is not the only entity that people tag—by better incorporating the multifaceted ways people describe *music* into recommenders, we are beginning to explore richer forms of user feedback applicable to *all* recommenders.

ACKNOWLEDGMENTS

We thank the study participants for their time and feedback. This work was made partly possible through the ERASMUS program.

REFERENCES

[1] Bertin-Mahieux, T., Eck, D., Maillet, F., and Lamere, P. 2008. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research.* 37, 2 (2008), 115–135.

[2] Bischoff, K., Firan, C.S., Nejdl, W., and Paiu, R. 2008. Can all tags be used for search? In *Proc. CIKM*, ACM, 193–202.

[3] Crampes, M., Villerd., J, Emery, A., and Ranwez, S. 2007. Automatic playlist composition in a dynamic music landscape. In *Proc. SADPI*, ACM, 15–20.

[4] Cunningham, S.J., Bainbridge, D., and Falconer, A. 2006. More of an art than a science: Supporting the creation of playlists and mixes. In *Proc. ISMIR,* 240–245.

[5] Divisi. http://csc.media.mit.edu/divisi (accessed April 2012)

[6] Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., and Pierce, L.G. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies*. 58, 6 (Jun. 2003), 697–718.

[7] Gemmis, M.D., Lops, P., Semeraro, G., and Basile, P. 2008. Integrating tags in a semantic content-based recommender. In *Proc. RecSys*, ACM, 163–170.

[8] Hansen, D.L., and Golbeck, J. 2009. Mixing it up: recommending collections of items. In *Proc. CHI*, ACM.

[9] Herlocker, J., Konstan, J.A., Borchers, A., and Riedl, J. 1999. An algorithmic framework for performing collaborative filtering. In *Proc. SIGIR*, 230–237.

[10] Inskip, C., MacFarlane, A., and Rafferty, P. 2010. Upbeat and quirky, with a bit of a build: Interpretive repertoires in creative music search. In *Proc. ISMIR*, 655–661.

[11] Kapoor, A., Lee, B., Tan, D., and Horvitz, E. 2010. Interactive optimization for steering machine classification. In *Proc. CHI*, ACM, 1343–1352.

[12] Knijnenburg, B.P., Reijmer, N.J.M., and Willemsen, M.C. 2011. Each to his own: How different users call for different interaction methods in recommender systems. In *Proc. RecSys*, ACM, 141–148.

[13] Kulesza, T., Stumpf, S., Burnett, M., and Kwan, I. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proc. CHI*, ACM.

[14] Kulesza, T., Stumpf, S., Wong, W.-K., Burnett, M., Perona, S., Ko, A., and Obsert, I. 2011. Why-oriented end-user debugging of naive Bayes text classification. *Transactions on Interactive Intelligent Systems*, 1, 1 (Oct. 2011).

[15] Last.fm Audioscrobbler API. http://www.last.fm/api (accessed June 2011).

[16] Law, E.L.M., von Ahn, L., Dannenberg, R., and Crawford, M. 2007. Tagatune: A game for music and sound annotation. In *Proc. ISMIR*, 631–634.

[17] Levy, M. and Sandler, M. 2008. Learning latent semantic models for music from social tags. *Journal of New Music Research,* 37, 2 (2008), 137–150.

[18] Liu, H., and Singh, P. 2004. ConceptNet—a practical commonsense reasoning tool-kit. *BT Technology Journal,* 22, 4 (2004), 211–226.

[19] Music genome project. http://www.pandora.com/mgp.shtml (accessed April 2012)

[20] Open mind common sense. http://openmind.media.mit.edu/ (accessed April 2012)

[21] Pachet, F. and Cazaly, D. 2000. A taxonomy of musical genres. In *Proc. Content-Based Multimedia Information Access*, 1238–1245.

[22] Pazzani, M.J., and Billsus, D. 2007 Content-based recommendation systems. *LNCS* 4321, 325–341.

[23] Rand, W.M. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 336 (1971), 846–850.

[24] Schafer, J.B., Frankowski, D., Herlocker, J., and Sen, S., 2007. Collaborative filtering recommender systems, *The Adaptive Web*, Springer Verlag.

[25] Sordo M., Gouyon F., and Sarmento L. 2010. A method for obtaining semantic facets of music tags. In *Proc. WOMRAD*.

[26] Sparling, E.I. and Sen, S. 2011. Rating: How difficult is it? In *Proc. RecSys*, ACM, 149–156.

[27] Specia, L. and Motta, E. 2007. Integrating folksonomies with the semantic web. In *Proc. ESWC*, 624–639.

[28] Speer, R., Havasi, C., and Lieberman, H. 2008. AnalogySpace: Reducing the dimensionality of common sense knowledge, In *Proc. AAAI*, 548–553.

[29] Stumpf, S. and Muscroft, S. 2011. When users generate music playlists — when words leave off, music begins? In *Proc. AdMIRe*, *ICME*, 1–6.

[30] Stumpf, S., Rajaram, V., Li, L., Burnett, M., Wong. W.-K., Dietterich, T., Sullivan, E., Drummond, R., Herlocker, J. 2009. Interacting meaningfully with machine learning systems: three experiments. *International Journal of Human-Computer Studies,* 67, 8 (2009), 639–662.

[31] Sukumaran, A., Vezich, S., McHugh, M., and Nass, C. 2011. Normative influences on thoughtful online participation. In *Proc. CHI* (2011), 3401–3410.

[32] Talbot, J., Lee, B., Tan, D., and Kapoor, A. 2009. EnsembleMatrix: Interactive visualization to support machine learning with multiple classifiers. In *Proc. CHI*, ACM, 1283–1292.

[33] Tang, J., Leung, H., Luo, Q, Chen D, Gong J. 2009. Towards ontology learning from folksonomies. In *Proc. IJCAI*.

[34] Tatli, I. and Birtürk, A. 2011. Using semantic relations in context-based music recommendations. In *Proc. WOMRAD*.