Exploiting Recommended Usage Metadata: Exploratory Analyses

Xiao Hu
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
xiaohu@uiuc.edu

J. Stephen Downie
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
jdownie@uiuc.edu

Andreas F. Ehmann
Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
aehmann@uiuc.edu

Abstract

In this paper, we conduct a series of exploratory analyses on the user-recommended usages of music as generated by 1,042 reviewers who have posted to www.epinions.com. Using hierarchical clustering methods on data derived from the co-occurrence analyses of usage and genre, usage and artist, and usage and album, we are able to conclude that further investigation of user-recommended usage metadata is warranted, especially with regard to its implications for future iterations of the Music Information Retrieval Evaluation eXchange (MIREX).

Keywords: user-recommended usage, music reviews, hierarchical clustering, MIREX, HUMIRS, user study

1. Introduction

Questions involving how music information is used have drawn growing attention in the MIR community. Studies have explored music use behaviors from different angles and by various approaches. For example, Bainbridge et al. [1] employed qualitative ethnographic methods to examine music seeking behaviors and goals as presented on the Google Answers website. Lee and Downie [2] applied survey methods to develop a set of descriptive statistics on information needs, uses and seeking behaviors amongst a group of 427 users. Data mining and natural language processing (NLP) techniques have also been applied on a variety of music related textual data sources to analyze how users organize and access music information. For example, Whitman and Lawrence [3], Baumann and Hummel [4], and Knees et al. [5] retrieved top ranked web pages related to a list of artists from search engines, applied text processing techniques to build vector space models representing the artists and then calculated artist similarities based on similarities of the vectors. Pachet et al. [6] used playlists from radio stations and compilation CD databases to conduct cluster analyses on music titles and artists. User-generated playlists have also been used to compute artist similarities which were subsequently validated against ground truth taken from the expert-derived artist lists compiled by the editors of www.allmusic.com [3].

Online music reviews, such as those found on www.epinions.com, because they are written by music users themselves, provide both users' opinions and an excellent source for ground truth data with regard to such things as genre labels, artist names, quality ratings, etc. For example, Hu et al. [7] developed a set of Naïve Bayesian models that could successfully predict both the user-assigned genre labels (78.9%) and the user-generated quality ratings (81.3%) based upon the examination of 2,160 epinions.com review texts. More recently Downie and Hu [8] undertook a frequent pattern analysis of the epinions.com review data and discovered a set of three-term descriptive patterns that were helpful in differentiating positive and negative reviews.

In this paper, we extend the previous work on the analysis of user-generated music reviews but now shift the analytic focus to explore the user-generated recommended *usages* of the music being reviewed. In each of the reviews presented on epinions.com, there is a field called "Great Music to Play While" where the reviewer provides a usage suggestion for the reviewed piece. Each review can be associated with at most one recommended usage. Epinions.com provides a ready-made list of 13 recommended usages prepared by the editors. By analyzing the application of these recommended usage labels using a range of statistical analyses, we aim to answer the following questions:

1. What are the relationships between usages and music genres?
2. What are the relationships between usages and music artists?
3. How are the usages related to each other based on their co-occurrences with genre, artist, and album titles? Can the usage classes be meaningfully grouped into broader categories (i.e., superclasses)?

1.1 Motivation

The present exploratory study is being conducted under the auspices of the Human Use of Music Information Retrieval Systems (HUMIRS) project. The goal of the HUMIRS project is the creation of standardized evaluation tasks and "query" documents, grounded in real-world user behaviors, for use in the Music Information Retrieval Evaluation eXchange (MIREX). The overarching goal of this study is the determination of whether (or not) user-generated "usage" classes warrant further examination as contributory elements in: (a) the definition of future MIREX contests; and/or, (b) the basic structure of the MIREX evaluation "query" documents.

2. Data Set

Table 1 shows the recommended usage categories applied to 5,772 review texts randomly downloaded from epionions.com in April 2005 along with the number of reviews suggesting a particular usage. Our analyses in this paper focus on the first 11 categories and ignore the last 2 categories where the data were too sparse for meaningful analyses (i.e., "With Family" and "Sleeping"). 180 reviews from each of the 11 usage categories were selected to form a data set of 1,980 reviews. Such even distribution across usage categories was made to eliminate possible analytic biases.

Table 1. Suggested usage categories and counts

Usage	Count	Usage	Count
Driving	1,349	Waking Up	271
Hanging With Friends	1,215	Going to Sleep	269
Listening	592	Cleaning the House	230
Romancing	492	At Work	188
Reading or Studying	447	With Family	35
Getting Ready to Go Out	378	Sleeping	15
Exercising	291	TOTAL	5,772

3. Co-occurrence Analyses

The review dataset we are studying contains information about the genre of the album, the artist who performs the music, the album title under review and the usage suggested by the reviewer. In this section we examine the relationship between (a) music genre and usage; (b) artist and usage; and, (c) album and usage.

3.1 Genre and Usage

3.1.1 Dependency Analysis

There are 12 genres covered by the 1,980 reviews in the dataset used in our experiments. These were the same 12 genres we examined in our previous work [7]. To determine the relationships between different genres and different usages, a vector, consisting of 23 binary variables (representing the 12 genre and 11 usage classes), was constructed for each review. These 1,980 vectors formed the input for a Pearson's chi-square dependency test [9] for all possible pairs of genres and usages. Table 2 reports the χ^2 and p values of only those tests that indicated strong dependencies between the corresponding genres and usages (statistically significant at the $p < 0.01$ level). Some of the results validate "common sense" about possible genre and usage relationships, such as "Jazz Instrument" music being significantly related to "Romancing." Some others, however, are interesting findings that might not have been well known beforehand, such as "Electronic" music being recommended for "Going to Sleep," and "Country" music for "Cleaning the House."

Table 2. Key dependencies between genres and usages

Genre	Usage	Pearson's χ^2	p value
Classical	Listening	37.61	< 0.001
Country	Cleaning the House	70.78	< 0.001
Electronic	Going to Sleep	29.13	< 0.001
Hard Core Punk	Waking Up	12.54	< 0.001
Jazz Instrument	Romancing	123.45	< 0.001
Pop Vocal	Romancing	49.88	< 0.001

3.1.2 Usage Clustering on Genre-Usage Co-occurrences

A thoughtful examination of the usages listed in Table 1 suggests that some of the usages are semantically related. For example, "Exercising" and "Cleaning the House" both denote some kind of physical activity. While it is tempting to have the authors themselves manually group the categories into superclasses, this approach is obviously not preferable because the interpretation of the semantics of the category labels can be subjective. To overcome author-based biases, we instead chose to take advantage of the associations between music genre and usage to see if it could assist us in generating meaningful superclasses upon which to base further explorations. Therefore, we clustered usages based on their co-occurrences with genres. A co-occurrence matrix was formed such that each cell of the matrix was the number of reviews with the genre and usage specified by the coordinates of the cell. This co-occurrence matrix then underwent a hierarchical clustering procedure using Euclidean distance and the complete link algorithm [10]. Figure 1 shows the clustering results.

We note here the grouping of the "Read/Study" and "Go to Sleep" usages that suggests a kind of *Passive* or *Relaxing* engagement with the music. This contrasts well with the more *Active* or *Stimulating* interactions evoked by the grouping of the "Ready to Go" and "Exercise" usages.

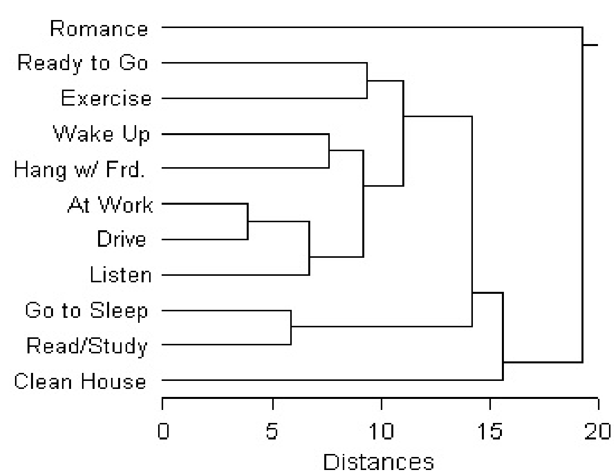

Figure 1. Usage clusters from genre-usage co-occurrences

3.2 Artist and Usage

3.2.1 Dependency Analysis

There are 897 unique artists in the dataset of 1,980 reviews. It is interesting to see how specific artists tend to be recommended for certain usage occasions. We explored the data to discover the strongest relationships between specific usage classes and specific artists. In order to verify that our observations were not purely random, a binomial exact test [11] for each of the artists and usages was conducted on those artists that were represented by at least 10 reviews. Table 3 presents the artist usage pairs whose relations tested to be significant at $p < 0.05$. For example, Black Sabbath's music is recommended most often for use "At Work." Similarly, a significant plurality of reviewers recommended listening to Nirvana when "Going to Sleep."

Table 3. Significant relations between artists and usages

Artist	Usage	*p* value
AFI	Waking Up	0.034
Black Sabbath	At Work	0.000
Celine Dion	Romancing	0.025
Dream Theater	Listening	0.019
Metallica	Waking Up	0.033
Nirvana	Going to Sleep	0.019

We suggest that artist-usage relationships warrant further examination on two fronts. First, work needs to be done to see if a case could be made that the music produced by these artists, in some way "exemplifies" or represents the "archetype" of the associated usage classes. For example, constructing an audio-based feature set of "Metallica" and "AFI" pieces for exploratory similarity experiments based on the identification of other "Waking Up" (or more generally, *Active/Stimulating*) music could prove fruitful. Second, notice in Figure 2 how each artist appears to have a unique recommended usage profile (i.e., various proportions of recommend usages). If one were able to garner enough usage data for each artist to construct his/her own usage profile, such data could contribute significantly to the building of artist-to-artist similarity analyses and recommender systems.

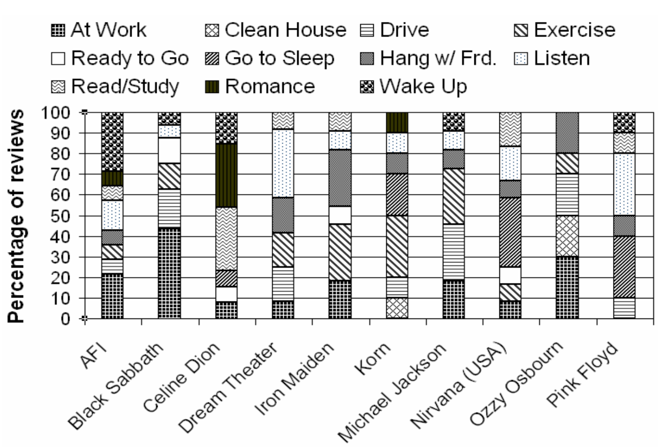

Figure 2. Usage distributions of 10 most-reviewed artists

3.2.2 Usage Clustering on Artist-Usage Co-occurrences

The facets of artist and usage can also form a co-occurrence matrix. Using the same clustering techniques discussed in 3.1.2, usage clusters were created from the artist-usage co-occurrences (Figure 3).

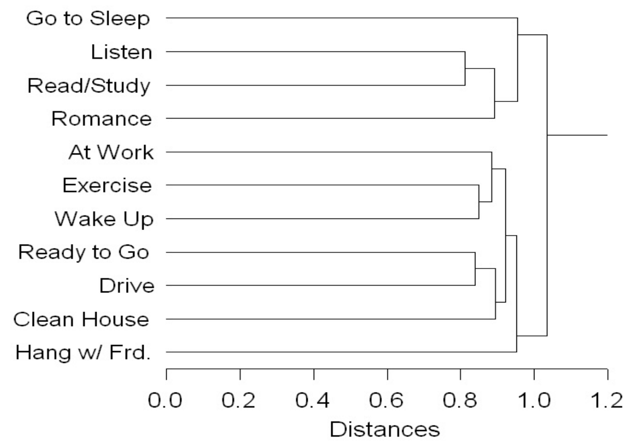

Figure 3. Usage clusters from artist-usage co-occurrences

Again we note in Figure 3 that the close proximity of the "Go to Sleep" and "Read/Study" usages suggests a *Passive/Relaxing* grouping. Similarly, the close proximity of the "Exercise" and "Waking Up" usages suggests an *Active/Stimulating* relationship with the music.

3.3 Album and Usage

3.3.1 Usage Clustering on Album-Usage Co-occurrences

Album titles are another facet that can be utilized as the hidden variable that connects different usages. There are 1,372 unique albums in this dataset, among which 366 albums have 2 to 7 reviews. A co-occurrence matrix of albums and recommended usages was constructed as before and was subjected to the aforementioned clustering procedures. The resulting cluster tree is shown in Figure 4.

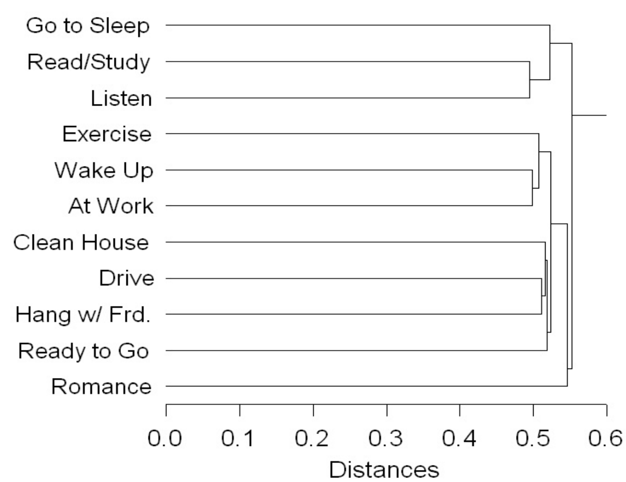

Figure 4. Usage clusters from album-usage co-occurrences

4. Discussion

Comparing Figures 1, 3, and 4, it is obvious that there are consistent general patterns to the clusters derived from three completely different data facets. Again, we note that "Go to Sleep" and "Read/Study" are always grouped together (i.e., *Passive/Relaxing*), while "Ready to Go," "Exercise," "Wake up," "Drive," "Hang out with friends," and "At Work" (i.e., *Active/Stimulating*) are consistently grouped together at various levels of proximity. This general consistency of the clustering methods' results across different data facets (genre, artist, and album) serves to reinforce the notion that the usage groups of *Active/Stimulating* and *Passive/Relaxing* are reasonable and that the groupings do indeed seem to suggest the presence of at least two usage superclasses.

However, there are still individual recommended usages that show discrepancies in groupings across experiments. One problematic example is "Listening": it is grouped together with different usages across the various experiments. This disagreement is not surprising because from a semantic point of view, "Listening" is an ambiguous usage, as all music activities involve listening. Although there should be a reason for a user to pick "Listening" among all the given choices, it is unclear whether the reason relates more to *Active/Stimulating* or *Passive/Relaxing* activities.

There are several limitations to the dataset that could play a role in masking the presence of other readily discernible and informative superclasses. First, the usage labels were created by the editors of epinons.com and thus end users did not have complete freedom to specify their usage recommendations but rather had to choose from the 13 options provided. Second, some of the options are semantically ambiguous such as "Listening" mentioned above. Third, it is reasonable to say that most of the usages may not have been interpreted consistently by different users. Fourth, and finally, the users were limited to only one usage choice, which thus prevents us from examining how individual users would have brought together related usage classes to form possible superclasses.

5. Conclusions and Future Work

In this exploratory study we set out to determine whether user-recommended usage information warranted further examination in the construction of MIREX tasks and/or "query" documents. The consistent manifestation, under a range of experimental conditions, of groupings of usage categories that form the two plausible superclasses of *Passive/Relaxing* and *Active/Simulating* leads us to conclude that the recommended usages specified by users reflect a meaningful source of user-generated metadata. As such, we firmly believe that further investigation is warranted so that we may better understand this particular type of user-generated metadata in order to better model real-world user behavior for MIREX.

In future work, further analysis on the relationships among genre, artist and usage is desirable, as is the exploration of new, large-scale, data sets. We are especially interested in determining whether other (and possibly more informative) usage superclasses can be uncovered. We also want to move toward examining the relationships between the recommended usages and the audio features of the music under review.

6. Acknowledgments

We thank the Andrew W. Mellon Foundation for their support. This project is also supported by the National Science Foundation under Grant No. NSF IIS-0327371.

References

[1] D. Bainbridge, S. J. Cunningham and J. S. Downie. "How People Describe Their Music Information Needs: A Grounded Theory Analysis of Music Queries. In *ISMIR 2003 Fourth Int. Conf. on Music Inf. Retr. Proc*., 2003, pp. 221–222.

[2] J. Lee, and J. S. Downie. "Survey of music information needs, uses, and seeking behaviours: Preliminary findings," *ISMIR 2004 Fifth Int. Conf. on Music Inf. Retr. Proc*., 2004, pp 441– 446.

[3] B. Whitman, and S. Lawrence. "Inferring Descriptions and Similarity for Music from Community Metadata," In *Proceedings of the 2002 International Computer Music Conference*, 2002, pp. 591–598.

[4] S. Baumann and O. Hummel. "Using Cultural Metadata for Artist Recommendation," in *Proceedings of the WedelMusic Conference*, 2003, pp. 138–141.

[5] P. Knees, E. Pampalk, and G. Widmer. "Artist Classification with Web-based Data," in *ISMIR 2004 Fifth Int. Conf. on Music Inf. Retr. Proc*., 2004, pp. 517–524.

[6] F. Pachet, G. Westerman, and D. Laigre. "Musical Data Mining for Electronic Music Distribution," In *Proceedings of the WedelMusic Conference*, 2001.

[7] X. Hu, J. S. Downie, K. West, and A. Ehmann. "Mining Music Reviews: Promising Preliminary Results." in *ISMIR 2005 Sixth Int. Conf. on Music Inf. Retr. Proc*., 2005, pp. 536-539.

[8] J. S. Downie and X. Hu. "Review Mining for Music Digital Libraries: Phrase II," In *Proceedings of the Sixth Joint Conference on Digital Libraries*, 2006, pp.196-197.

[9] A. Agresti. *An Introduction to Categorical Data Analysis*, Wiley, New York 1996.

[10] A. Jain, M. Murty and P. Flynn. "Data Clustering: A Review," *ACM Computing Surveys*, Vol 31, No. 3, pp.264-323, 1999.

[11] M. Buntinas, and G. M. Funk. *Statistics for the Sciences*, Brooks/Cole/Duxbury, 2005.