MIR TASK AND EVALUATION TECHNIQUES

First author
Affiliation1
author1@ismir.edu

Second author
Retain these fake authors in submission to preserve the formatting

Third author
Affiliation3
author3@ismir.edu

ABSTRACT

Existing tasks in MIREX have traditionally focused on low-level MIR tasks working with flat (usually DSP-only) ground-truth. These evaluation techniques, however, can not evaluate the increasing number of algorithms that utilize relational data and are not currently utilizing the state of the art in evaluating ranked or ordered output. This paper summarizes the state of the art in evaluating relational ground-truth. These components are then synthesized into novel evaluation techniques that are then applied to 14 concrete music document retrieval tasks, demonstrating how these evaluation techniques can be applied in a practical context.

1. INTRODUCTION

ISMIR split from the JCDL conference in 2000 as an independent conference specializing in the creation of systems that perform music specific tasks especially the retrieval of musical entities from a collection or database. Since then, papers such as Tzanetakis' Marsyas paper [16] and Goto's database [8] provided a rich set of tools and techniques and, in the process, defined both the tasks of MIR and the ground truth to compare against. The IMERSAL project [7] took this a step further, providing an evaluation platform that now defines what is MIR.

As more systems are created for music document retrieval, there is a need to standardize tasks and the techniques for evaluating their performance. Several different disciplines have focused on different aspects of this evaluation technique. A summary of the different evaluation techniques is presented in Section 2.

From the beginning of MIREX [5], obtaining quality ground-truth has been a perennial problem. With the arrival of relational ground-truth, not only is acquisition of ground-truth problematic, the sampling of ground-truth sources also need to be synchronized with each other. Section 4 presents novel approaches for addressing these problems.

The use of multiple data sources in a single algorithm requires new evaluation techniques. Section 5 describes the components of novel evaluation techniques for relational ground-truth. Included is the definition for a novel metric, Serendipity.

Finally, Section 6 presents a set of standardized task descriptions spanning music document retrieval, many of which are novel problem descriptions. These descriptions include all the details needed to implement a MIREX task that can be conducted by the IMERSAL [6, 7] without unduly straining existing resources.

2. RELATED WORK

The earliest attempt to standardize MIR evaluation is Downie's 2002 paper [6], defining the problems for MIR in general without describing solutions for any particular task in detail. Downie and Cunninghams's paper [5] hinted at the different MIR document retrieval tasks possible, but did not enumerate them. Downie et al.'s followup paper on MIREX in 2005 [7] gives an overview of the tasks that have evolved (quite different from the initial goals of the conference). While much of the discussion moved to the MIREX wiki and mailing lists, two more papers on specific tasks were published in 2008 [4, 11] Afterwards, the discussions moved to MIREX wiki and email lists.

In non-ISMIR disciplines, collaborative filtering [9] provides a wealth of research on evaluating ordered or numeric output. In addition, relational machine learning [12] have evaluated the statistical problems utilizing cross-validation with relational data.

3. NOTATION

While the notation of relational analysis may be more familiar to readers, graph theory notation is used for its simplicity describing complex relationships between data sources. Throughout this paper, a single table entry of a data source is referred to as a 'node' The name of a data source is referred to as a 'mode'. if a relationship exists between two table entries (such as a foreign key), that relationship is a 'link'. The name of a foreign key is defined as a 'relation'. Graphs are defined as a set of nodes and links.

4. GROUND TRUTH ACQUISITION

Ground truth acquisition can be subdivided into two parts: what data to include and how to sample these data sources to create ground-truth for an experiment. Acquiring this data is now much more easily accomplished utilizing crawling of available on-line data sources while utilizing sampling techniques from sociology provide both representa-

Figure 1. Description of the Semantic Web in 2008 [15]

tive subsets of these sources as well as data synchronization.

4.1 Data Sources

Fortunately, there have been a number of online sources in the past few years providing source data that greatly reduces the severity of the ground truth problem. The OMRAS project [15] has provided a great deal of support for linking these data sources together. Likewise, GData provides an even wider knowledge base by establishing relationships between other data sources. LastFM provides REST access to much of their data. MySpace has an online dataset of popular music now mirrored in FOAF formats [15].

These resources are in REST format (XML documents acquired by HTTP request). While most utilize the Freind-of-a-friend data format (with extensions) [15], LastFM utilizes a different syntax.

4.1.1 MySpace Data

MySpace is mirrored by the MyData web site. Included are FOAF-syntax references to all publicly distributed mp3s on MySpace. Links to LastFM artists are provided via artist names or indirectly via a lookup of an artist or track MBID through the MusicBrainz services.

4.1.2 GData

The GData protocols link existing blog and profile pages. In addition, it provides protocols for linking to Fickr and YouTube resources.

4.1.3 DBpedia

This resource bundles Wikipedia in an XML syntax. While it does not include DSP-related data, it has detailed descriptions of artists, albums, and tracks that are organized into a hierarchal structure of categories. This knowledge base is then embedded in a general knowledge base structured in a graph form. Access to this data is either via online services or by downloadable database snapshots. Links to and from DBpedia require URL construction.

4.1.4 MusicBrainz

MusicBrainz provides identity hashes that provide clean metadata that indexes nearly all published music. The hashes can be used as keys for LastFM, eliminating ambiguity with artists or albums of the same name.

4.1.5 Other On-line Data

Numerous resources including LiveJournal profiles, Google blogs, online ontologies, and numerous other online data sources are now provided as XML feeds. With the exception of LiveJournal profiles (FOAF format), these references typically require URL construction to link them with other data sources. see figure 1 for a description of available data sources as documented by the OMRAS project.

4.1.6 LastFM Data

LastFM web services provide access to nearly all data collected excepting personal profile information. Artist, album, and track information are all indexed by MusicBrainz hashes that provide links between LastFM data and the rest of the semantic.

4.1.7 Playlists

The Art of the Mix [1] website provides a collection of user-generated playlists that, coupled with the metadata provided at the site, provide a set of ground truth and queries for a number of different tasks. This data set can be acquired via web services, however choosing queries and ground truth from this data set must still be done manually.

4.2 Sampling Techniques

The most traditional sampling technique for information retrieval is random sampling. This approach is appropriate if data is flat, as it implicitly assumes that any particular element chosen is independent of any other element. In relational contexts, it has two major drawbacks: it distorts the underlying graph structure of the data and the precision and recall are dependant of the degree of dependency between the folds (a factor that is difficult to correct in evaluation.)

In response to these drawbacks, the social network analysis sub-discipline of sociology developed technique called 'snowballing' [17]. In this technique, a start node is chosen and all actors within a pre-determined number of links from this starting place are added to the data set. This technique (implemented in Graph-RAT) automates the process of synchronizing sampling of data sources in a way that guarantees synchronization of the data.

Even after these techniques are applied, there are difficulties in sychronizing ground truth with available playlists. In practice, this will either require utilizing ground-truth play-lists that provide an incomplete list of valid songs, or utilize the Evaluatron system [11] to rate the extent an algorithm has satisified the constraints implied by the query play-lists.

[1] `http://www.artofthemix.org`

5. EVALUATION TECHNIQUES

Evaluation techniques can be subdivided into three different components - ground truth acquisition, segmentation, and evaluation metrics. Section 4 describes ground truth acquisition.

5.1 Cross-Validation Techniques

In order to prevent over-fitting, four different forms of segmenting ground truth into testing and training data have been proposed across IR. While the literature frequently uses different terminology, each of these cross-validation techniques can be described in terms of randomly removing graph nodes, graph links, or utilizing distinct subgraphs of relational data. Each of these techniques are implemented in the Graph-RAT software package [2].

5.1.1 Node-Based Cross-Validation

Node-based cross-validation can be accomplished by random removal of nodes from a source graph during training, implicitly removing all links to this node. During testing, all other nodes of the same type are removed instead. Oscar Celma's recent dissertation [3] provides a detailed description of this approach, utilizing node-based cross-validation across either track nodes or user nodes. This particular approach can not evaluate collaborative filtering systems as the information utilized to produce results are removed by the evaluation technique. In addition, the use of random sampling assumes that the nodes removed are independent—an assumption violated by relational ground-truth. This approach is the standard approach for MIR systems.

5.1.2 Link-Based Cross-Validation

Link-based cross-validation separates links of a particular relation into testing and training sets. Algorithms are evaluated on whether they can recreate the witheld links. The most obvious example includes randomly removing links between users and tracks. This approach typically contains most of the dependencies between users and tracks intact—allowing evaluation of collaborative filtering systems. However, it is randomly determined if all links of a given track or user are removed, the independence assumption of cross-validation is violated more severely than actor-based cross-validation. Despite this, it does allow collaborative filtering systems to evaluated in a more robust manner than not withholding testing data.

5.1.3 Graph-Based Cross-Validation

This approach is typical in relational machine learning evaluation. Typically, hand generated training and evaluation data sets are used that guarantees independence of training and evaluation data. Jensen [12] proposes an algorithm for automating this process for segmenting graphs with minimum disruption of the underlying link structure.

[2] http://graph-rat.sf.net

5.1.4 No Training-Testing Subdivisions

This approach evaluates with full knowledge of the training data in advance. Whole easy to implement, this technique has serious problems over-estimating the performance of algorithms. It is the primary evaluation technique of the collaborative filtering sub-discipline [9].

5.2 Evaluation Metrics

The techniques for evaluating an IR task depend greatly on the kind of output they generate. Collaborative filtering [9] in particular utilizes a number of metrics that capture other properties of an algorithm's output beyond binary present-not present. These metrics can be divided into three categories. Correlation requires an unordered numeric data, precision and recall require set-based data. Ranked list require ordered output that is either an ordered weighted list or a specific play-list order with internal constraints.

5.2.1 Correlation Metrics

These metrics evaluate a play-list by correlation or other form of vector-based similarity metric between derived and ground-truth results. These are typically evaluate the return of numeric values for each returned item without any concern for the order of the results. Mean Absolute Error [2], Pearson Correlation [10], and Spearmans Correlation [9] provide this kind of evaluation.

In addition, the author proposes a new metric, called Serendipity. It evaluates the originality of an algorithm's output, addressing problems with produdcing novel recommendations as described in [9]. It is defined as

$$Serendipity = \frac{\sum_{i=1}^{n} trackfrequency_i}{\sum_{j=1}^{n} trackfrequency_j} \qquad (1)$$

where i is the ith member of the list of most popular tracks and j is the jth member of the derived list. This metric provides a numeric value indicating how atypical the prediction is.

5.2.2 Precision and Recall Metrics

These metrics evaluate binary, unordered results that are the standard of machine learning including precision, recall, and F-measure. When evaluating the results of multiple queries, audio-cover [4] has used two internally-created metrics: mean reciprical rank and average performance. Mean Average Performance [4, 14] was also utilized in 2007.

5.2.3 Ranked List Metrics

These metrics evaluate play-lists where the order is important. It differs from correlation based techniques as the order in which results are presented alter the outcome even if the entries are equal in magnitude. This is especially important in play-list generation scenarios where part of the query are constraints across track boundaries. Matching tempo between sequential tracks is one example this kind of constraint. Half-life [2], NDPM Measure [18], Kendall

Tau's Correlation [9] and matching techniques such as Hamming Distance and Dynamic Time Warp are the metrics in this category.

5.3 Significance Tests

While a number of significance tests described in the literature, ISMIR tasks have standardized [4,11] on a combination of Friedman's ANOVA [1] followed by Tukey-Kramer Honestly Significantly Different analysis [1].

6. TASK DEFINITIONS

Herlocker et al. [9] defines document retrieval in terms of a set or purposes: Annotation in context, find good items, recommend sequence, just browsing, and find credible recommender. In contrast to this approach, each task in music entity retrieval (along with its implied purpose) is described seperately, allowing more detailed recommendations for how to evaluate this kind of task, implicitly defining the purpose of the task. Each of these tasks can be categorized by 4 binary attributes producing a different combination of data sources, query, cross-validation technique, and evaluation metric to evaluate algorithms accurately. These tasks can be further subdivided into subtasks via additional attributes that do not require structural changes to evaluate an algorithm's performance on these subtasks.

6.1 Binary features of a musical information need

The different musical item information needs can be categorized by a combination of a number of different citeria. These criteria are whether results are restricted to new music only, dynamic versus static data, open versus restricted queries, and personalized or general results. Each combination requires different ground-truth, training data, query structure, and evaluation metrics.

6.1.1 New Music Restriction

This criteria is whether results are restricted to music that is not previously known. In the general case, new music is music that has no social information, only DSP data. In the personal data, this can be relaxed to only music that is not already in a user's playlist information.

6.1.2 Dynamic Versus Static Data

LastFM data can be acquired for ewach week, not just in aggregate. Dynamic data utilizes time-dependent data while static data uses the data in aggregate only.

6.1.3 Open Versus Restricted Queries

Restricted queries are looking for specific information (such as all files in a particular genre). These restrictions are to be presented as a set of playlists demonstrating the restriction(s) by example. Open queries return all documents, ordered by specific criteria such as a Top 40 radio station or a personalized radio station.

All tasks utilizing restricted queries utilize custom playlists for ground truth. This requires either hand selecting relevant play-lists or creating custom play-lists. While much can be automated utilizing key word searches on the metadata of existing play-lists, the accuracy of the evaluation of these tasks are typically restricted by the extent that the ground truth covers all valid results.

6.1.4 Personalized Versus General

Some algorithms are indpendent of a particular user such as browsing online music listings. Others include the profile of a user in the algorithm.

Beyond these criteria, there exists additional considerations such as placement- and context-sensitive evaluation, play-list length, and DSP only evaluation. These factors can be accommodated by differing evaluation techniques or algorithm restrictions independent of the structure of the task.

6.2 Comprehensive Description of Tasks

Each combination of the binary features describes a separate information need utilized in a particular kind of MIR. All tasks will utilize Friedman's ANOVA [1] followed by Tukey-Kramer Honestly Significantly Different analysis [1]. with

6.2.1 Personalized radio

The personalized radio task utilizes a personal profile with no restrictions, static data, and no requirement that the music be new. It requires the static version of LastFM user data along with the knowledge base. In order to evaluate collaborative filtering approaches, a link-based cross-validation technique across the User-¿Track relation is needed. A correlation metric defining likes/dislikes or a ranked metric is most appropriate for the final evaluation. Pandora and LastFM are examples of this task.

6.2.2 Top 40

The top 40 task utilizes only the knowledge base. It is a sanity check of all algorithms in that the final outcome is a trivial-to-acquire ground truth—the current Top 40 chart. A graph-based cross-validation technique should be used as collaborative filtering does not work without personalized data and this is the most statistically valid partitioning of relational data. Also, it should be evaluated with a ranked metric.

6.2.3 Tag Radio

The tag radio task track utilizes only the knowledge base. The query is a set of constraints provided indirectly via play-list examples. The result is an ordered list representing the best ordering of musical entities in a search result. Like Top 40, this should use graph-based cross-validation with a ranked metric. This task emncompasses most of the existing MIREX tasks such as genre classification, mood classification, and cover song identification.

6.2.4 Personalized Tag Radio

The personalized tag radio task utilizes LastFM profile data and the knowledge base. The query is a set of constraints

and a user profile. Even though the technique utilizes personalized data, existing collaborative handling system can not handle constraints, so graph-based cross-validation with either a ranked metric or string matching metric is the most appropriate evaluation technique. Another application of algorithms in this track is perrsonalized ordering of music in a hierarchal presentation of all music.

6.2.5 Personalized Radio with Dynamic Data

The persoanalized radio with dynamic data track utilizes LastFM weekly top tracks, LastFM profile data, and the knowledge Base. It provides a personalized radio station that auto-corrects as changes occur in a persons tastes over time. It is designed to be a more difficult task then personalized radio alone—attempting to predict specifically the best songs for the next week, not just best songs in general. No cross-validation should be performed and the final week of the LastFM data set withheld as the testing set. The final result should be evaluated either with a correlation or ranked metric.

6.2.6 Personalized Tag Radio with Dynamic Data

The personalized tag radio with dynamic data task utilizes LastFM weekly tracks, LastFM profile data, and the knowledge base. The query is the profile and a set of constraints. Like personalized Tag Radio, this should use graph-based cross-validation with either a string matching or ranked metric. This can also be used for a time-sensitive personalized ordering of music in browsing applications.

6.2.7 Personalized New Music Radio

The personalised new music task utilizes LastFM profile data and the knowledge base. Utilizing actor-based cross-validation on track actors, all non-DSP data from the music is removed. For any particular user, only music not already in the user's profile is included in the evaluation. No further cross-validation is performed as the lack of social data makes the testing set fully independent of the training set already. The results are evaluated either with a correlation or ranked metric against the base LastFM profile data. This is equivelant to a new-to-me personalized radio station.

6.2.8 New Top 40

The new top 40 task utilizes only the knowledge base. Like the personalized version, actor-based cross-validation on tracks is performed where all non-DSP data is removed. The evaluation should the difference between the sum of all weekly tracks in the last two weeks of data to create a ranked list, evaluated with a ranked metric. This is essentially a 'hit predictor' track.

6.2.9 New Music Tag Radio

The new music tag radio task is similar to its non-new-music equivelant except only new music is considered. It uses the same cross-validation technique as personalized new music radio, however, the ground truth is the total of the number of occurances in constraint play-lists that are in the current testing set.

6.2.10 Personalized New Music Tag Radio

The personalized new music tag radio task utilizes LastFM profile data and the knowledge base. It utilizes the same cross-validation techniques as personalized new radio. Ground truth, however, is the profile weightings for each track in the current fold.

6.2.11 Personalized New Music Search with Dynamic Data

The personalized new music search with dynamic data task utilizes LastFM weekly data, LastFM profile data, and the knowledge base. The same cross-validation techniques should be applied as for the Personalized New Music Radio track. The evaluation should be with a ranked metric against the last week of data in the lastFM weekly data.

6.2.12 New Top 40 with Dynamic Data

The new top 40 with dynamic data task utilizes the LastFM weekly data in aggregate and the knowledge base. It is a more difficult version of the New Top 40 track where prediction should only be of new music listened to in the next week. The actor-based cross validation used for Personalized Top 40 should be used here as well, but the ground truth should be the sum of all weekly data that is in the current fold. The metric should be a ranked metric.

6.2.13 New Tag Radio with Dynamic Data

The new tag radio with dynamic data task is a more challenging version of the personalized tag radio with dynamic data track utilizing lastFM weekly data, LastFM profile data, and the knowledge base. It should utilize the same actor-based cross-validation as the personalized new music track. It requires special constraint play-lists that only utilize tracks that appear in a particular week of listening. Furthermore, users should be filtered from the dataset if they do not listen to any of the resulting tracks. The ground truth for any particular fold is the sum of all play-list entries that are in the current fold. The results should be evaluated with a ranked metric.

6.2.14 New Music Tag Radio with Dynamic Data

The new music tag radio with dynamic data task is a more challenging version of the tag radio with dynamic data track utilizing the LastFM weekly data in aggregate and the knowledge base. It should utilize same actor-based cross-validation techniques as the new personalized radio track. The ground truth is the same as personalized new tag radio with dynamic data except that the play counts are aggregated over all users. The results should be evaluated with a ranked metric.

7. CONCLUSION

In this paper, a summary of available techniques for evaluation of relational ground-truth is presented. In addition, a new metric, Serendipity, is presented. Techniques for sampling and synchronizing on-line ground-truth sources are

also presented. These techniques are then applied to 14 concrete music document retrieval tasks.

8. FUTURE WORK

The collection and annotation of non-DSP data requires more development for restricting manual ground-truth collection exclusively to query construction. Furthermore, additional work is needed to integrate these tools into existing evaluation frameworks such as NEMA and the IMERSAL lab.

9. ACKNOWLEDGMENTS

Omitted for blind review

10. REFERENCES

[1] M. L. Berenson, M. Goldstein, and D. Levine. *Intermediate Statistical Methods and Applications: A Computer package approach*. Prentice-Hall, 1983.

[2] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Uncertainty in Artificial Intelligence*, San Francisco, USA, 1998. Morgan Kaufmann.

[3] Oscar Celma. *Title*. PhD thesis, Univeritie of Pampfrea, 2008.

[4] J. Stephen Downie, Mert Bay, Andreas F. Ehmann, and M. Cameron Jones. Audio cover song identification: Mirex 2006-2007 results and analysis. *Internation Conference on Music Information Retrieval*, 2008.

[5] J. Stephen Downie and Sally Jo Cunningham. Toward a theory of music information retrieval queries: System design implications. *International Conference on Music Information Retrieval*, 2002.

[6] Stephen J. Downie. Towards the scientific evaluation of music retrieval systems. *ISMIR*, 2003.

[7] Stephen J. Downie, Kris West, Andreas Ehmann, and Emmanuel Vincent. The 2005 music information retrieval exchange (mirex 2005): Premilinary overview. *ISMIR*, 2005.

[8] Masataka Goto and Hiroki Hashiguchi. Rwc music database: Popular, classical, and jazz music databases. *International Conference on Music Information Retrieval*, 2002.

[9] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering systems. *ACM Transactions on Information Systems*, 22(1):5–53, January 2004.

[10] W. Hill, L. Stead, M. Rosenstein, and G. W. Furnas. Recommending and evaluating choices in a virtual community of use. In *Confernce on Human Factors in Computing Systems*, New York, USA, 1995. ACM.

[11] Xiao Hu, J. Stephen Downie, Cyril Laurier, Mert Bay, and Andreas F. Ehmann. The 2007 mirex audio mood classification task: Lessons learned. *International Conference on Music Information Retrieval*, 2008.

[12] David Jensen. Knowledge discovery through induction with randomization testing. In G. Piatetsky-Shapiro, editor, *Proceedings of the 1991 Knowledge Discovery in Databases Workshop*, pages 148–159, Menlo Park, 1991. American Association for Artificial Intelligence.

[13] Daniel McEnnis and Sally Jo Cunnigham. Sociology and music recommendation. *International Conference on Music Information Retrieval*, 2007.

[14] M. G. Di Nunzio, N. Ferro, T. Mandl, and C. Peters. Clef 2007: Ad hoc track overview. *Working Notes for the CLEF Workshop*, 2007.

[15] Yves Raimond and Mark Sandler. A web of musical information. *International Conference on Music Information Retrieval*, 2008.

[16] George Tzanetakis and Perry Cook. Audio information retrieval (air) tools. *International Conference on Music Information Retrieval*, 2000.

[17] Stanley Wasserman and Katherine Faust. *Social Network Analysis*. Structural Analysis in the Social Sciences. Cambridge University Press, Cambridge UK, 2 edition, 2004.

[18] Y. Yao. Measuring retrieval effectiveness based on user preference of documents. *ASIS*, pages 133–145, 1995.