Validation of Music Metadata via Game with a Purpose

Peter Dulačka
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies,
Slovak University of Technology,
Bratislava, Slovak Republic
dulacka@gmail.com

Jakub Šimko
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies,
Slovak University of Technology,
Bratislava, Slovak Republic
jsimko@fiit.stuba.sk

Mária Bieliková
Institute of Informatics and Software Engineering
Faculty of Informatics and Information Technologies,
Slovak University of Technology,
Bratislava, Slovak Republic
bielik@fiit.stuba.sk

ABSTRACT
Quantity of music metadata on the Web is sufficient, music recommendation and online repository systems are proof of it. However, it became a real challenge to keep quality of these metadata at reasonable level as the cost of manual validation is too high and current automatic approaches are inaccurate. In this paper we present a game with a purpose called City Lights – a music metadata validation approach which lowers the cost of human computation and makes the validation fun. Our goal is to get rid of incorrect user-submitted music tags or tags not usable at global scale. We describe the game principles and evaluate the game results.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]: Content Analysis and Indexing; K.8 [**Personal Computing**]: Games

General Terms
Design, Experimentation

Keywords
game with a purpose, human computing, multimedia, music information retrieval, metadata validation

1. INTRODUCTION
In order to provide best search results or recommendations of multimedia resources, it is essential to assign correct metadata to them. The general groups of approaches to do this are: (1) automatic acquisition, (2) crowdsourcing or (3) expert involvement. There are negative effects with every approach (high cost – money or human hours, wrong metadata generation); therefore researchers combine them to gain the best possible results. However, even after this, there always are some metadata assigned incorrectly. The *task of validation of already existing metadata* then becomes a relevant issue.

Our work deals specifically with music tags. Music information retrieval (MIR) is still not evolved enough to gain mandatory music metadata by automatic means [6]. Combined with crowdsourcing methods (especially social tagging or online bookmarking) it provides higher quality information, yet also human workers introduce noisy metadata due to heterogeneous use cases, subjective views, etc.

Music metadata can be divided into three categories:

- Objective, acquired automatically (e.g. rhythm, melody)
- Objective, crowdsourced (e.g. author, song title, etc.)
- Subjective, crowdsourced (e.g. mood, quality, time and place to listen)

All groups also comprise noisy tag samples, especially the last one, which is strongly influenced by subjective opinions of contributing users. This noise lowers the quality of crowdsourced metadata and should be removed from datasets through the validation process. The validation of (music) metadata is in its nature a different process than their creation and is usually not a task that users would do willingly.

To motivate people to perform this task, we propose a game with a purpose (GWAP) called *City Lights*. The game provides fun to its players via a *competition* and *music exploration*. Meanwhile it uses the game logs to *validate existing music tags*. The basic task for the player in the game is to guess, which set of tags was originally assigned to the music track that is currently being played (the tag set is presented with other existing sets from different tracks). Upon the decisions and the confidence of the player, our method afterwards infers the validity of featured tags.

In this paper we describe our game and its task: validation heuristics. We evaluate the game, reporting on the live experiment we performed with it.

2. RELATED WORK
GWAPs in general are type of games being used to solve problems which machines cannot solve accurately, but humans can solve them without hard effort (i.e. HITs – human intelligence tasks). The first game with a purpose called ESP Game was created by von Ahn and Dabbish [9]. The game focused on image labeling and opened new grounds in human computation. After its success (both in fields of gained metadata and player satisfaction) von Ahn came with object position obtaining game [10] which used annotations from ESP Game. He also devised music related game Tagatune where he focused on retrieving new music annotations - a multiplayer game where player had to create annotations according to what she heard and decide whether she listened to the same song as her randomly chosen co-player [3].

Besides von Ahn's projects, approaches with different game mechanisms were introduced. Šimko et. al. presented a single-

I-SEMANTICS 2012, 8th Int. Conf. on Semantic Systems, Sept. 5-7, 2012, Graz, Austria

Figure 1.Game interface: music player (left), game board (centre), annotation container (right), game log (bottom), window for annotation marking (on top).

player game for creating folksonomy-like network relationships [6] which used number of search engine query results to evaluate player's actions. Morton et. al. presented Moodswings, which obtained mood of songs [5] by mapping mood into 2D graph and was able to capture different moods throughout the song. The game showed that even more abstract games can generate usable and accurate metadata as well. The most successful game related to music information retrieval is HerdIt [1]. In the game players listen to short song previews and have to choose correct word describing the song. This method turned out as player friendly as no typing or creativity was necessary. Both acquisition methods (tag typing and option picking) were later used in Listen Game [7] where players had to first choose one of words describing song (passive acquisition) and then type the most accurate song characteristics (active acquisition). None of the existing GWAPs focuses primarily on validation of metadata and use validation just as a part of annotation creation process.

Common problem with most existing games with a purpose is a cold-start and inability to overcome it. It is caused mostly by insufficient number of players and data for acceptable gameplay experience, unattractive level design, cheating possibilities or poorly designed scoring [4, 7].

Most of the presented games acquire and validate metadata at the same time. This is achieved by real-time, two- (or more-) player game where players have to agree on round's input or output. This however brings cheating possibilities: when players know each other, they may arrange the game round in order to achieve the highest score. To prevent this, games try to match players randomly, but there is no guarantee that matched players will not know each other (especially in games with low player base).

Problems described above point to most important issues to focus on while designing game with a purpose: (1) attractive level design and simple gameplay, (2) proper and motivating scoring and (3) building up a player base. Metadata validation is an important challenge [8], so the annotations can be properly used in projects such as [2].

3. MUSIC METADATA VALIDATION

Our game-based method focuses on the validation of existing music tags fetched from the Web. Player is presented with several sets of real music track annotations (a bag of already existing tags, acquired by both automatic and combined approaches) each set related to different song. Player then hears a part of a music track and decides, which of the given sets relates to track he is listening to. Then, through a betting mechanism, player expresses how sure she is about her choice (he bets a certain number of points) and gets rewarded for good and punished for bad guesses.

The basic premise of the tag validation through such scheme is that player's behavior implicitly signals the discerning power and ultimately the "correctness" of the tags assigned to tracks present in the game. If she guesses the correct set, some of the tags present in that set *probably* represents the playing track. If, on the other hand, she guesses a wrong set, then it is more likely that the correct set (the one she was supposed to identify) contains less related tags to the playing track and the wrongly selected one contains something more characteristic to it.

If a particular track occurs multiple times in the game for multiple players, the correctness probabilities of individual tag-track relationships get summed and uncertain and wrong tags may be filtered out. To boost the whole process, the method also offers an explicit feedback option for players to rule out tags they consider wrong (in return, these tags would not display to them in the future games so they stand chances for higher point gain). The betting mechanism also helps: the player confidence is in a direct relation with weight of the implicit feedback acquired.

3.1 City Lights Game: the Player View

We have realized the metadata validation method within the casual game called City Lights[1] (see screenshot in the Figure 1). Player travels the graph of streets (edges) and crossroads (nodes) of a city and on each crossroad makes a decision about which way to take: she listens to music and has to pick correct metadata set.

[1] Accessible at: http://bit.ly/city-lights

The interface of the game consists of music player on the left, the city plan in the middle (with circles representing the crossroads), annotation set container at the right and a game log at the bottom. The actual position (crossroad) of player and possible directions to choose are always highlighted. After pointing on one of the directions, a set of annotations appears in annotation set container. Player can pause or rewind the song and is not limited by time. The level design is straight and goes as follows:

1. Player is given a game board with highlighted initial crossroad and possible directions. Each node of the game board is related to exactly one song and its set of annotations.
2. Music player starts playback and player is allowed to explore annotations related with possible direction crossroads. Based only upon these annotations, she has to decide which of the available crossroads contains annotations related to the song. Initially, the annotation sets contain a fixed number of tags (in our experiments, we used 5). For a small point fee, the player may disclose a few more tags in the set to her aid.
3. When attempt is made, player chooses a bet height (effectively a confidence expression about the choice).
4. After making a correct decision, she eventually marks incorrect tags and proceeds to the next song. Music is still being played in the background to make the decision easier.
5. Game ends when user reaches final crossroad.

3.2 Metadata Processing

The source for the metadata validation process is the game log, which comprises:

- The sequence of player's actions (e.g. crossroad decisions, "more tag" requests, "incorrect tag" exclusions).
- The setup for each game (what tags were displayed, how many choices the player had on each crossroad, what tags were not displayed but still assigned to a particular track in the source corpus).

Having these data at hand, we created several heuristics for estimating the correctness of tags assigned to music track (from the point of their general usability). All of these heuristics manipulate with so called *support* value – an expression of a probability that a tag is correctly or incorrectly assigned. This value is initially set to zero and is iteratively modified by heuristics triggered by players' actions. When *support* value reaches a positive or negative threshold, the tag is excluded from the process (and the game) as either confirmed or rejected.

If the attempt (tag bag guess) of the player was *incorrect*, it gives us two important messages: (1) provided annotations for the track being played may not be accurate enough (so we implicitly decrease their *support*) or/and (2) annotations for different song are accidentally better describing, than provided ones (we increase their *support* in case they are also present in the track set). We also give the player option to explicitly mark annotations which persuaded her to select the wrong set (small score reward is given, this action is optional). Marked annotations then become possible annotations for track being played, even though they are not present in the source corpus (explicit *support* increase).

If the attempt was *correct*, the *support* for tags displayed in the selected set should be increased (implicit *support* increase). However, if the correct attempt was preceded by previous incorrect attempt(s), this increase is lesser (in our game setup, a marginal after two incorrect attempts). As in the previous case, the player has the option to rule out confusing tags, this time those not describing the track. These tags then receive explicit *support* decrease instead (while other, presumably correct tags, have it increased, splitting the rejected tag original *support* increase).

The key feature of the game is, that tag sets are continuously changed so individual tag strongly influencing player decisions can be recognized: if a particular tag perfectly describes the track its *support* gets increased each time, while other, not-so-good or wrong tags also receive decreases. In order to process every tag as fast as possible, simple rule is being applied: The more extreme (more distanced from its initial value) is the *support* of a tag for particular track, the bigger is the chance of its appearance in next game round so the decision about its relevance is quicker.

3.3 General Game Design Issues

In order to provide fun and player satisfaction, we used numerous approaches described in [3]. Player enjoyment is provided by high-score lists and other social connections, challenges (e.g. bonus points for passing a level without a fault or possibility to post the result to social network) and randomness of songs played in each game. However in order to make players like the game, we cannot play completely random songs - we provide songs from domain she entered at registration. Thanks to single-player game design there are no cold-start problems and cheating possibilities, though still present, are significantly lowered.

The most important part of design is proper scoring. We decided not to release scoring formulas to players; however they are familiar with get-points-for-everything scheme. By giving points even for actions not connected with our purpose it is more likely that players will return in the future. In the beginning of a game round the player is given certain initial number of points she may bet, based on number of tracks in particular game. Score of every action is counted by number of incorrect attempts made on particular node and the level of certainty that player has chosen. If player chooses incorrectly, the number of points based on chosen confidence is subtracted from her score.

If score drops to zero, the game is over. There are conditions which prevent players to make no-risk attempts, so the fear of losing points makes them consider the option they choose better.

4. EVALUTATION

We have performed an experimental evaluation of the tag validation capabilities of our approach. We implemented the game as a web application, let the players play, collected the game logs and performed tag validation computations. Then, we computed the method's accuracy using a comparison to apriori created golden truth data set of tagged music tracks prepared by experts.

Hypothesis. Our game-based method is able to identify correct (objective and distinctive) metadata (tags) assigned to music tracks, drawn from larger sets containing both correct and incorrect metadata.

Data. We used 100 music tracks. Their annotations were fetched from public LastFM[2] database. For each track we acquired 40 top tags (from the original LastFM ranking), removed 10 top to preserve reasonable game difficulty and further considered them as of equal rank. Track previews were being played using 7Digital[3] library. To create the golden data set, we invited 3 judges – a people experienced in music domain (musician, music historian, collector) and asked them to identify which of the tags

[2] http://www.lastfm.com
[3] http://www.7digital.com

they perceive correct. They first worked independently and then had to reach consensus about each tag. The expert evaluation rendered 44% tags to be correctly assigned to the music tracks.

Participants. In total 78 players (Web and social network users) participated in the experiment. We considered no prior knowledge about their demographical characteristics.

Environment and context. The experiment was conducted in an uncontrolled environment. The potential players learned about the game via social networks or e-mail we sent them. The participation was purely voluntary and uncontrolled.

Process. The game playing lasted 10 days. In total, 875 games were played with total of 4 933 puzzles ("crossroads" with a decision) solved. Overall 1 492 tags appeared in the game at least one time. Each tag received averagely 17.75 implicit feedback actions (changes of *support* as a consequence of "crossroad decision making") and 5.29 explicit feedback actions (explicit player inclusions or exclusions of tags for a particular track).

Results. In the experiment, we tested various combinations of parameters of our method, to find out the best possible setup (with fixed positive threshold = 5 and negative threshold = -5). By testing multiple combinations of attributes values we identified: *implicit tag support increase* = 0.2, *implicit tag support decrease* = 0.3, *explicit tag support increase* = 0.9, *explicit tag support decrease* = 0.9. The output parameters which we aimed to optimize were:

- *false negative ratio* denoting the percentage of correct tags that method has rejected, the most important one: optimized to 0%, i.e. no correct tags were rejected,
- *validation ratio* denoting the percentage of tags about which the method stated some result: optimized to 49%, mainly due to the fact that method has not received enough feedback on many tags,
- *false positive ratio:* optimized to 38%.

Out of the processed tags 729 were identified as correct, 39 as incorrect and 724 received not enough player feedback to be evaluated by our method. The method correctly evaluated 66% of tags which we find promising, though the method was unsure in many cases, mostly due to lack of collected feedback. More importantly, the method is able to filter out tags which are certainly not correct, which is its primary task and also correctly confirms tags at decent rate.

Upon first inspection of the filtered metadata, we could see that our method got rid of annotations such as: *elotmbgmegamixx, test, nice, favorite, good lyrics, fab, etc.* These are either very subjective annotations not usable at global scale or complete nonsense. On the other hand method approved tags as: *female vocalists, love, british, singer-songwriter, pop rock, etc.* Some of them are subjective, but usable at global scale and some of them are objective and should be validated without any trouble.

During the experiments, the game and the ladder competition really challenged and motivated many players. However, some also reported that their main reason to play the game further was the exploration of new music or simply an enjoyment of music. Even after the end of experimental period, several players continued to play, regarding the game as a good procrastination tool. This stresses the game's potential to be widely used in a casual gaming scenario.

5. CONCLUSION AND FUTURE WORK

Our experiments has shown that the game *CityLights* is able to harness players' brain capacity to successfully perform the task of cleansing the bag of tags related to music tracks and that it provide promising degree of accuracy. Meanwhile the single-player nature of the game imposes no cold-start problems as it is in cases of many other games with a purpose.

As for our future work, we plan to improve player scoring and tag evaluating to achieve better user experience and even more accurate results. We plan to consider decision time as a factor in evaluating annotations and create a model where each player has different impact on the evaluation according to her experience. We have ambition to add multiplayer mode, where two players could play the same game in different time so they can directly compare their results. There are also open possibilities for measuring player's performance in particular music domains (e.g. rock, pop, metal) and exploiting this knowledge when composing the in-game puzzles for him.

6. ACKNOWLEDGEMENT

This work was partially supported by the grants VG1/0971/11/2011-2014, VG1/0675/11/2011-2014 and APVV-0208-10.

The authors wish to thank colleagues from the Institute of Informatics and Software Engineering and all students (members of PeWe group, pewe.fiit.stuba.sk) for their invaluable contribution to the experiment presented in this paper.

7. REFERENCES

[1] Barrington, L., O'Malley, D., Turnbull, D., Lanckriet, G.: User-centered design of a social game to tag music. *Proc. of the ACM SIGKDD Workshop on Human Computation - HCOMP '09*, 2009, p. 7.

[2] Dittmar, C., Großmann, H., Grollmisch, S., Lukashevich, H., Abeßer, J.: Two Applied Research Projects in Music Inf. Retrieval at Fraunhofer IDMT. 2011, pp. 259-272.

[3] Law, E. L. M., von Ahn, L., Dannenberg, R. B., Crawford, M.: Tagatune: A game for music and sound annotation. 2007, pp. 361–364.

[4] Mandel, M. I., Ellis, D. P. W.: A web-based game for collecting music metadata. *Journal of New Music Research*. 2008, vol. 37, no. 2, pp. 151–165

[5] Morton, B. G., Speck, J. A., Schmidt, E. M.: Improving music emotion labeling using human computation. *Proc. of the ACM SIGKDD Workshop on Human Computation.* 2010, pp. 45-48.

[6] Šimko, J., Tvarožek, M., Bieliková, M.: Little search game: term network acquisition via a human computation game. *Proc. of the 22nd ACM conf. on Hypertext and hypermedia*. 2011, pp. 57-61.

[7] Turnbull, D., Liu, R., Barrington, L., Lanckriet, G.: A game-based approach for collecting semantic annotations of music. *Int. Sym. on Music Information Retrieval.* 2007, pp. 2-5.

[8] Turnbull, D.: Automatic Music Annotation. Department of Computer Science and Engineering, University of California, San Diego, CA. Research Exam. 2005.

[9] von Ahn, L., Dabbish, L.: Designing games with a purpose. *Comm. of the ACM*. Aug. 2008, vol. 51, no. 8, p. 57.

[10] von Ahn, L., Liu, R., and Blum, M.: Peekaboom: A Game for locating objects in images. *In Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM Press, New York, 2006, pp. 55–64