Music Artist Tag Propagation with Wikipedia

Luís Sarmento  
LIACC/FEUP, Univ. do Porto  
Rua Dr. Roberto Frias, s/n  
Porto, Portugal  
las@fe.up.pt

Fabien Gouyon  
INESC Porto  
Rua Dr. Roberto Frias, 378  
Porto, Portugal  
fgouyon@inescporto.pt

Eugénio Oliveira  
LIACC/FEUP, Univ. do Porto  
Rua Dr. Roberto Frias, s/n  
Porto, Portugal  
eco@fe.up.pt

ABSTRACT
In this paper we present a method for automatically assigning tags to music artists in the Web 2.0 radio Last.fm. Starting with the reference list of Last.fm user-defined tags, our method mines Wikipedia music artist abstracts for new tag candidates. Tag candidates are ranked using an heuristic weighting function. We evaluate the top ranked tag suggestion for about 31,000 artists by (i) performing automatic evaluation using diachronic Last.fm data, and (ii) by performing manual evaluation on a sample of artists. Our method shows promising results regarding the accurate propagation of artist tags.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—Information filtering; H.3.5 [Information Storage and Retrieval]: On-line Information Services—Web-based services; H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing

General Terms
Algorithms, Experimentation, Performance

Keywords
Social media analysis, Discovery-driven mining, Information extraction and summarization, Last.fm, Autotagging

1. INTRODUCTION
One of the most interesting advances of the Web 2.0 is the possibility for users themselves to add and edit meta-information about content by assigning descriptive tags to media items. Such social tagging process leads to the emergence of meaningful textual descriptions of web content that can be extremely helpful in information retrieval tasks, especially when automated content analysis is still not accurate enough (e.g. video or music). However, social tagging mechanisms tend to lead to unbalanced tag distributions: while popular items are tagged many times by many users, less popular item will probably only receive a few tags, or even none at all. Popular items are abundantly described by tags while less popular items might not have enough tags—both in quantity and diversity—to have meaningful, and stable descriptions. Some authors refer to this as the “Cold start problem” [5], [10]. If tag information is used for retrieval, then less popular items will probably be retrieved less times, degenerating in a retrieval starvation effect. If tags are used for music recommendation, the most tagged artists end up biasing recommendations [3].

In this paper we focus on a specific Web 2.0 site, http://www.Last.fm, which allows users to tag both artists and songs. Users associate different types of tags to artists [5], tags can be related to: music genre (e.g. “acid jazz”), locale (e.g. “japan”), artist/band structure or instrumentation (e.g. “duo”), personal experiences (“seen live”), opinion (“weird”) and all sorts of miscellaneous tags (e.g. “Evrovision”). Figure 1 shows the distribution on the number of tags per artist for a universe of 583,497 artists, using data taken from Last.fm webservice in February 2008 (see Section 4). The vast majority (almost 80%) of the artists has 5 or less tags and 47% has not been tagged at all.

Figure 1: Artist/Tags distribution

In this paper, we tackle the problem of unbalanced tag distribution by using Wikipedia abstracts to extract, rank and finally assign an additional tag to a significant subset of Last.fm artists.

2. RELATED WORK
There are a number of proposals in the literature for dealing with unbalanced tag distributions of music items. For instance, following image tagging research [12], some authors propose to add information to music items via entertaining games [6], [7], [11]. Another line of work focuses on prop-
agating tags from popular artists (i.e. tagged frequently) onto other, less frequently tagged artists. A technique to do this is to project artists into a similarity space, and to propagate tags to close neighbors in that space. Such a similarity space can be constructed by content-based similarity computation [9], [2] (i.e. assuming that music items that “sound alike” should be tagged alike — this has been referred to as autotagging), or via collaborative filtering or co-occurrence analysis [4] (i.e. assuming that music items that are commonly found together, e.g. in different users’ playlists, or in webpages [8], should be tagged alike). For a complete review, and list of applications of music social tagging, we refer to [5], data and bibliographic links can also be found on http://SocialMusicResearch.org.

3. TAG PROPAGATION METHOD
Let $B_{last}(a_i)$ be the “bag” of user-defined tags found in February 2008 on Last.fm for artist $a_i$. Our approach consists in using semi-structured third-party information sources to perform tag propagation on Last.fm artists. Namely, we mine Wikipedia abstracts for $a_i$ to find relevant tags to be added to $B_{last}(a_i)$.

For example, the Wikipedia abstract for the band “!!!” is: “!!! (pronounced as chk chk chk, to simulate mouth-clicking sounds) is an American dance-punk band that formed in autumn 1996 from the former band members of The Yah Mos, Black Liquorice and Popesmashers.”. This small text passage contains information that could be used for tagging “!!!”, namely [American] and [dance-punk]. From Last.fm page dedicated to “!!!” we can in fact confirm that users as-signed both [american] and [dance-punk] to the band. More important cases are those for which there are no user-defined tags in Last.fm. For example, at the date of writing, there are no Last.fm tags for the band “Brixx”, while the corresponding Wikipedia abstract describes the band as follows: Brixx was a Danish pop group which represented Denmark in the Eurovision Song Contest 1982, in which it sang “Video.”. This contains valuable tagging information as e.g. [Danish pop], [Denmark], [Eurovision Song Contest 1982].

Because tags can be extremely diverse in nature we opted for considering only those tags that have already been assigned to some artist in Last.fm. This can be achieved by building a Tag Dictionary, $T_{last}$, from all the tags used in Last.fm, and matching only elements that are part of that dictionary against Wikipedia abstracts. The tag propagation procedure for a given Last.fm artist $a_i$ can thus be performed using the following procedure:

1. Check if there is a Wikipedia article for artist $a_i$. This is done by matching the artist name with the article title and ensuring that certain music-related words (e.g., “singer”, “band”, “music”, “artist”, “composer”, “group”) are found in the abstract to reduce probability of processing irrelevant / ambiguous names;
2. If a Wikipedia article is found, then try to match tags from $T_{last}$ on the article abstract. This will create $T(a_i)$ containing all tags matched.
3. Remove from $T(a_i)$ all tags already in $B_{last}(a_i)$ and rank each remaining tag according to a relevance function (see Section 5).

4. DATA PREPARATION
Some of the data related to Last.fm radio (artist, users, etc.) is freely available through a dedicated web-service API\(^1\). For a period of about a week (from January 30 to February 4 2008) we consulted Last.fm web-service to obtain a local copy of data concerning artists and their user-defined tags. We obtained basic information for 583,497 artists (name, “popularity” index within Last.fm community) and information regarding 2,774,068 tag attributions. On average we found 4.76 tags per artist, but many artists do not have any tag assigned (see Figure 1). There are 208,565 distinct tags, a surprisingly high number. The 10 most commonly used tags are: “seen live” (54,660 artists), “rock” (41,854), “electronic” (33,108), “indie” (27,913), “alternative” (25,401), “pop” (24,010), “punk” (20,555), “electronic” (18,781), “metal” (17,419) and “experimental” (16,680).

Part of Wikipedia’s content has been converted into tabular format by the DBpedia project [1], allowing a simple access to certain parts of the content (e.g. infoboxes) without the need for performing complex parsing operations. In our work, instead of directly consulting Wikipedia articles, we used the short abstract data provided by DBpedia, which contains abstracts (1-3 sentences) for 2,491,442 entities/concepts identified by Wikipedia page title (the data we used was downloaded on October 20 2008).

5. EXPERIMENTAL SETUP
Our Tag Dictionary $T_{last}$ is composed by 182,556 tags: 71,875 with 1 word, 77,643 with two words and 33,038 with 3 words (all tags were converted to their low-case representation to avoid duplication derived from case variation). We ignored longer tags (4+ words) to optimize the matching procedure. For each tag $t_j$ in $T_{last}$ we computed $\#_{last}(t_j)$, the number of artists in Last.fm tagged with it by users. This statistic reflects the importance of tag $t_j$ in Last.fm user-defined tag follosomy. From the initial set of 583,497 artists it was possible to match 36,947 artist names with Wikipedia article titles. 987 of the artists names were considered ambiguous in Wikipedia, so we only kept the matches for 35,960 artists. Of these, only 3 Wikipedia abstracts did not match any single tag from $T_{last}$. This was a surprisingly low number but the explanation lies in the fact that there are many very frequent common words among user-defined tags, such as “a”, “for”, “with”, “is”, “he”. Since we have not performed any tag filtering on $T_{last}$, practically every abstract matched at least one element in $T_{last}$. However, only 30,114 tags of the 182,556 tags in $T_{last}$ were matched. For each of the 30,114 tags matched, we computed $\#_{wiki}(t_j)$, the number of Wikipedia abstracts which matched the tag $t_j$. Table 1 shows some illustrative examples of $\#_{wiki}(t_j)$ for several tags.

Tags were ranked according to the following weighting function:

$$w(t_j) = \frac{(n_{word}(t_j))^2 \cdot \#_{last}(t_j)}{\#_{wiki}(t_j)} \quad (1)$$

with $n_{word}(t_j)$ being the number of words of tag $t_j$ (one, two or three words). With this weighting function we seek to (i) demote tags that have been matched with many Wikipedia

\(^1\)http://www.audioscrobbler.net
\(^2\)http://dbpedia.org/
abstracts (e.g., “a”, “and”, “is”, “the”, “in”, “of”), since they have a very high probability of being noisy; (ii) promote tags which we know have already been assigned by users to many Last.fm artists, since this means they are relevant within Last.fm tag folksonomy; (iii) boost the relevance of relatively long tags (2 and 3 words) both because they are naturally more informative and have less chances of being noisy. Using this ranking function, the top 5 weighted tags are “seen live” (w = 218,640), “female vocalists” (w = 14,906), “drum n bass” (w = 11,124), “brutal death metal” (w = 6,951) and “folk metal” (w = 6,192). In the list of ranked tags for each artist ai, all tags with w(tj) < 0.25 were removed to avoid noisy assignments (albeit excluding some tags such as “band”, “rockband”, see e.g. Table 1). Following these steps, our method could propagate one new specific tag from Tlast onto 30,933 artists.

<table>
<thead>
<tr>
<th>matched tag tj</th>
<th># unique(tj)</th>
<th>w(tj)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>24,480</td>
<td>0.002</td>
</tr>
<tr>
<td>band</td>
<td>7,831</td>
<td>0.066</td>
</tr>
<tr>
<td>american</td>
<td>3,177</td>
<td>2.088</td>
</tr>
<tr>
<td>rock band</td>
<td>2503</td>
<td>0.247</td>
</tr>
<tr>
<td>singer</td>
<td>1,133</td>
<td>0.829</td>
</tr>
<tr>
<td>new york</td>
<td>517</td>
<td>11.520</td>
</tr>
<tr>
<td>country music</td>
<td>471</td>
<td>0.110</td>
</tr>
<tr>
<td>heavy metal band</td>
<td>228</td>
<td>0.039</td>
</tr>
<tr>
<td>classically trained</td>
<td>36</td>
<td>0.333</td>
</tr>
<tr>
<td>italian baroque</td>
<td>10</td>
<td>4.400</td>
</tr>
<tr>
<td>traditional instruments</td>
<td>8</td>
<td>1.500</td>
</tr>
<tr>
<td>swedish black metal</td>
<td>4</td>
<td>222.75</td>
</tr>
<tr>
<td>electro-indie</td>
<td>2</td>
<td>30.500</td>
</tr>
<tr>
<td>warehouse Rives</td>
<td>1</td>
<td>4.000</td>
</tr>
</tbody>
</table>

Table 1: Examples of tags, corresponding number of Wikipedia abstracts matched and weights computed by our ranking function.

5.1 Evaluation

We performed both automatic and manual evaluation on the best ranked tag suggestion only, sug_last(ai). Automatic evaluation consisted in comparing talast(ai) with the new tags actually assigned by Last.fm users to artist ai, between February 2008 and the date of writing (recall that Blast(ai) data was obtained in February 2008). For the 30,933 artists for which our method assigned a new tag, we queried Last.fm web-service to obtain current tag information. We found new user-defined for 23,310 artists (75.4% of 30,933), each having 15.4 new tags on average. Let us call this set of artists AS1 (for “Artist Set 1”). The remaining 7,623 artists (not further tagged since February 2008) will be called AS2. For artists ai in AS1, the set of new user-defined tags (i.e. those tags that were not in Blast(ai) but that users assigned to ai since February 2008) will be named Blast_new(ai) (“blast” = “new user-defined tags”).

For each artist in AS1, we computed the precision measure P Blast(ai). P Blast(ai) is 1 iff talast(ai) \in Blast_new(ai). This measure gives us an indication on whether our system can replicate the tagging behavior of Last.fm users during a 10 month period. We also defined the following, more permissive, yet informative precision measures: P Blast(ai) and P Blast_new(ai).

We performed manual evaluation for a random sample of 2% of the 7,623 artists in AS2. Using the information available in Wikipedia and in Last.fm artist pages, we manually computed the precision figure P manual(i) = 1 iff talast(ai) relates to (i) a possible music genre for the artist, or (ii) a specific style/attitude of the artist, or (iii) a geographic location relevant to artist biography, or (iv) relevant relations of the artist with other musical items (other artists, as e.g. former bands, record labels, etc.). P manual(ai) = 0 otherwise, i.e. incorrectly extracted, incomplete, ambiguous, irrelevant or uninformative tags were considered incorrect.

6. RESULTS AND ANALYSIS

Results of automatic evaluation on AS1 and manual evaluation on AS2 are presented in Table 2.

<table>
<thead>
<tr>
<th>AS1</th>
<th># artists (out of 23,310)</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>P Blast</td>
<td>1</td>
<td>1,793</td>
</tr>
<tr>
<td>P Blast</td>
<td>1</td>
<td>1,141</td>
</tr>
<tr>
<td>P manual</td>
<td>1</td>
<td>1,512</td>
</tr>
<tr>
<td>P Blast</td>
<td>1</td>
<td>823</td>
</tr>
<tr>
<td>P Blast</td>
<td>1</td>
<td>893</td>
</tr>
<tr>
<td>P Blast</td>
<td>1</td>
<td>6,162</td>
</tr>
</tbody>
</table>

2% AS2 # artists (out of 157) %
| P manual | 1 | 88 | 56.1% |

Table 2: Results of automatic evaluation for artist set AS1 (23,310 artists) and of manual evaluation for a random 2% sample of AS2 (i.e. 2% of 7,623 artists, i.e. 157 artists).

There is a relatively large difference in results of P manual and P Blast. Thus, we performed additional manual evaluation on AS1, to test whether such difference is due to differences between the two artist sets, or due to the very restrictive nature of the automatic evaluation procedure we proposed. By manual evaluating a %1 random sample of AS1 (i.e. 239 artists) we found 153 relevant talast(ai) suggestions, corresponding to a precision P Blast = 63.6%. This value, more in line with P Blast, confirms that our automatic evaluation procedure is in fact too restrictive.

A deeper analysis of the differences between automatic and manual evaluation unveiled many situations (i.e. 100 out
of 239, i.e. 42%) where tag propagation does add novel and relevant information about artist but the 5 performance measures fail to score it accordingly, because users tagging behaviour since February 2008 has been different (but not necessarily incompatible). For example, there are several cases where $t_{a_i}^{\text{ext}}$ refers to record labels (e.g. [universal records] for the band “Denver Harbor”, [infectious records] for “The D4”), or to additional activities of the artist (e.g. [project runway] for artist “Heidi Klum”, reflecting presence of the artist in a TV series). In other cases the additional tags assigned by users are actually irrelevant while $t_{a_i}^{\text{aut}}$ is correct (e.g. [hip hop] for “Gloria Velez”).

Error analysis revealed that there are two main causes of error: (i) incomplete tag extraction in 64 out of 239 cases (26%), and (ii) incorrect matching of Wikipedia page due to ambiguity in names (in 20 out of 239 cases, 8%). For instance artist “Ella Koon” was assigned the tag [french], when the relevant tag would be [french polynesia]. However, the tag [french polynesia] does not exist in Last.fm tag folksonomy, so only the known part of it (i.e [french] was extracted. Another example is the suggestion of [outstanding] to artist “Heinrich Wilhelm Ernst”, while the relevant tag would have been [outstanding violinist]. “Oil on Canvas” is a rather obscure band listed in Last.fm but is listed in Wikipedia as a live album by the British band “Japan”. Our simple disambiguation mechanism based on frequent music-related keywords, while relatively efficient in avoiding ambiguous names from other domains, is not able to avoid these ambiguous cases inside the music domain.

It is interesting to note that some tags considered correct by our automatic evaluation procedure seem to be relatively uninformative. These include both very frequently used (and thus highly ranked) tags such as “new”, “music” or “best”, as well as relatively obscure, vague or even noisy tags such as “pablo”, “sven”, “oc”, “e”, which end up being promoted by our ranking function because they are included in only a small number of Wikipedia abstract. All these tags are, nevertheless, part of Last.fm folksonomy. Other borderline cases are those of “redundant” tags. For example when the tag [charlotte perrelli] is assigned to artist “Charlotte Perrelli”, and [cowie] to “Chris Cowie”. These are valid tags (i.e. used by Last.fm users), but we may wonder whether they are really informative in the context of these artists.

7. CONCLUSIONS AND FUTURE WORK

We showed relatively encouraging results for the propagation of music artist user-defined tags by using third-party knowledge sources. So far, this has been an exploratory work, whose main aim was to test the feasibility of our approach to the “cold-start” problem. We found many lines for future work. First, we wish to expand the possibilities of propagating tags by (i) matching the entire content of the Wikipedia articles (not just abstracts) and (ii) by exploiting link structure on Wikipedia and Last.fm sites. Second, we will try to reduce the propagation of non-informative tags (that are nevertheless part of the Last.fm tag folksonomy) by improving the relevance ranking. We plan to include information regarding artist popularity in the weighting formula. Statistics taken from large external corpora might also help demoting irrelevant and non-informative tags. Third, we plan to mine user comments in Last.fm site for information that might help us differentiating between “consensual” and “subjective” tags (e.g., [the best]), and include such distinction in the ranking mechanism in order to demote the latter. Finally, since Last.fm tag folksonomy does not include all relevant concepts required for tagging many artists, we plan to expand it with new tag candidates discovered on external knowledge sources (e.g. geographic gazetters or the list of Wikipedia concepts). This would then allow us to suggest tags that have not yet been proposed by any Last.fm user but that are nevertheless useful for describing artists (e.g. locations or other musically-interesting concepts).

8. ACKNOWLEDGMENTS

This work was partially supported by grant SFRH/BD/23590/2005 from FCT

9. REFERENCES


