TOWARDS IMPROVED MUSIC RECOMMENDATION: USING BLOGS AND MICRO-BLOGS

1. Abstract

With the explosive growth of the World Wide Web and the rise of social media, new approaches in Music Recommendation evolve. The current study investigates how blogs and micro-blogs can improve the perceived quality of music recommendation. A literature review and expert interviews are conducted to identify important topics regarding (micro-) blogs and Music Recommendation. Subsequently, the prototype Songdice is built and tested in a user-evaluation. Songdice uses music blogs to recommend songs and rationalize those recommendations. Our results show that (micro-) blogs can improve the perceived quality of recommendations by creating trust, using personalization and exploiting the quality of music in the long tail. Additional research is required to determine the most effective way to use information from blogs and micro-blogs. Our research explores a new area in music recommendation literature and provides a starting point for further research concerning the combination of (micro-) blogs and music recommendation.

2. Keywords

Blogs; Music Recommendation; Music Information Retrieval; Twitter

3. Introduction

The explosive growth of the World Wide Web caused an enormous amount of multimedia creations. The huge volumes make it impossible for users to filter through the multimedia in order to find what they are looking for. One specific type of multimedia which is available in large quantities and varieties on the World Wide Web is music.

Several approaches to filter music have been researched and are used in recommendation engines. The general approaches are social (collaborative) filtering and content-based filtering (Basu et al., 1998). In recent studies, social media is linked to music discovery and recommendation. Laplante (2010) emphasizes the importance of recommendations from people one knows and trusts in music discovery. She proposes the inclusion of social networking tools in Music Information Retrieval (MIR) systems. Schedl (2010) has shown that Twitter posts provide a valuable data source for music information research, particularly in similarity estimation tasks. Zangerle et al. (2012) are the first to research the exploitation of tweets for music recommendation. To our best knowledge, no research is conducted on using blogs for music recommendation in the same way.

The aggregation of music blogs is used in several online services. *Shuffler*[i] and *The Hype Machine*[ii] use music blogs to feed their radio channel. However, they don't include a recommendation engine based on music weblogs. This paper will discuss how blogs and micro-blogs can improve the perceived quality of music recommendation. We thereby fill a gap in current research and provide a starting point for future research concerning the

combination of (micro-) blogs and music recommendation. This could eventually lead to a more qualitative recommendation process.

The following sections describe the used methodology, summarize the literature review, outline the results of expert interviews and user questionnaires, and describe and evaluate the prototype. We conclude with a summary of our findings and specific directions for further research.

4. Methodology

Our research starts with a review of the current literature. This review focuses on literature regarding recommendation engines, their approaches, evaluation processes and success, but also on blogs and micro-blogs.

From the literature review, important topics in music recommendation are selected. These topics are used to create a semi-structured interview for potential or current users of music-recommender systems. In addition, the topics help to define questions for several expert and user interviews. These interviews complemented the reviewed literature with personal insights.

The interviews are analyzed and combined with the literature. This analysis serves as input for the prototype requirements, in which technical and time-restriction are taken into account. Subsequently, a prototype that implements the requirements is designed and developed.

After developing the prototype, the user-evaluation is performed. A questionnaire is combined with the prototype to evaluate its quality. Hereby, the perceived quality of recommendations is measured. These results are used to draw a conclusion, propose improvements to the prototype and discuss areas for further research.

Figure 1 gives a visualization of the methodology of this study.

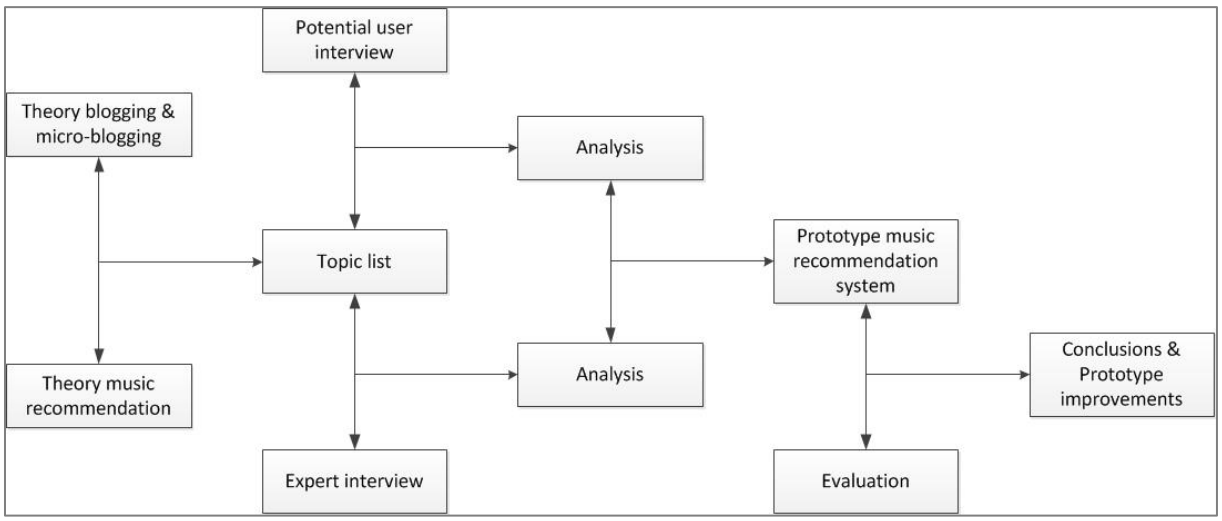


Figure 1, a visualization of the methodology.

5. Related work

This section presents a literature review on relevant research in Music Recommendation, music sharing through blogs and micro-blogs, improvements of perceived quality, combination of information and evaluation of recommendation systems.

Music Recommendation

A number of approaches to filter information have been researched and are used in recommendation engines. The two general approaches are social filtering and content-based filtering (Basu, Hirsh & Cohen, 1998).

- Social filtering, also known as collaborative filtering, is based on a virtual community in which users give ratings to a specified item. Based on these ratings, correlations are found with the ratings of other users for the same item. In this way, items that are rated positively by similar users can be recommended.
- Content-based filtering is based on the content of an item. Using a sample of user preferences, items with corresponding content can be recommended to the user.

Both approaches have their shortcomings (Balabanovic & Shoham, 1997). For example, social filtering suffers from the 'cold-start problem'. When items are new, they are not rated by any user and can thus not be recommended to other users. In addition, users with an unusual taste will not get any useful recommendations, because there are no similar users. Content-based filtering has some disadvantages too. Not always can relevant features be extracted from an item and recommended items are often very similar to liked items, which causes over-specialization. A mixture of both approaches is thus preferred and proposed by Balabanovic & Shoham.

Friend of a Friend

Celma (2010) presents a system called *FOAFing the music*. Its goal is "to recommend, to discover and to explore music content; based on user profiling, context-based information, and content-based descriptions". The context-based information describes contextual information of items (e.g. information from weblogs, reviews). The author makes use of 'Friend of a Friend' descriptions to build a user profile. These descriptions are concerned with the interests and relationships of a user. This hybrid approach thus makes use of three separate approaches.

Tags

With the rise of social networking sites such as *Facebook*, *Youtube*, *MySpace* and *Last.fm*, a new approach evolved, recommendation based on user-generated tags (Li et al., 2008). Firan et al. (2007) show that tag-based user profiles can provide better music recommendations on Last.fm than track-based profiles do. Guy et al. (2010) use a hybrid approach of both related people and related tags in their recommender system (which will be outlined in a later section). This approach achieves better results than the approaches that are based on only one of the aspects. Literature thus suggests that tags are of added value in recommender systems.

Micro-blogs

One of the latest trends in music recommendation seems to be the use of micro-blogs. Schedl (2010) has shown that Twitter posts provide a valuable data source for music information research, particularly in similarity estimation tasks. Zangerle et al. (2012) are the first to research the exploitation of tweets for music recommendation. Tweets with music-related keywords are crawled. In order to build a recommendation approach, songs which were

tweeted by the same user are stored in combination with the amount of users that tweeted both songs. In this way, triples in the form of (*t1*, *t2*, *c*) are created, where *t1* and *t2* are two tracks which are tweeted by *c* users. Finally, recommendation candidates are ranked, based on the count value *c*. This count value can be seen as the popularity of a certain track, compared to the tracks of a user's tweet-stream. Although the approach has to be enhanced to be usable in real-world recommendation environments, results are promising. However, Zangerle et al. mention limitations in Twitter's API, resulting in a sparse collection of tweets.

A shift from general approaches towards more social media-based approaches can be noticed from the literature review. Two recent trends involve the use of tweets and the use of tags in music recommender systems. Since results suggest or sometimes even show good performance, these approaches could be used in future applications.

Music sharing through blogs and micro-blogs

Blogs

Weblogs (blogs) are defined as "frequently modified web pages in which dated entries are listed in reverse chronological sequence" by Herring et al. (2004). According to the authors, blogs are seen as alternative sources of news and public opinion, environments for knowledge sharing and vehicles for self-expression and self-empowerment. These different 'roles' can all be related to music sharing. They can also provide advantages in music recommendation. First, an alternative source can present music that a lot of people do not know, i.e. the recommendations are relatively unknown songs. Second, the aspect of self-expression causes that a blogger posts about his diverse taste. One blog does not only contain similar songs, but also very different songs, which can bring variety in recommendation.

A myriad of weblogs concerned with music exists. Several online services (e.g. *Shuffler*[i] and *The Hype Machine*[ii]) use these music blogs to form a radio channel. The Hype Machine searches their collection of blogs in order to find mp3 files. Shuffler also takes other sources of music on blogs into account, e.g. YouTube videos and *Soundcloud* links. Blogs do not only share music file, but also information or reviews on artists, albums and songs (Celma, 2010).

Micro-blogs

Micro-blogging is an even more recent phenomenon on the World Wide Web. It gives users the opportunity to share what is on their minds in a very short message. *Twitter* is the most popular micro-blogging website, being the eighth most visited site in the world.[iii] In March 2012, 140 million people were actively using the site and 340 million tweets were sent per day. The service shows to be a rich source for information retrieval.

Twitter is often used to share opinions on artists, albums or songs (Schedl & Hauger, 2012). In addition, many audio players have a feature to automatically post the song a user is listening to on Twitter. The generated tweets usually contain keywords like *#nowplaying* or *listeningto* (Zangerle et al., 2012). Data from Twitter can be retrieved on an individual, city, country or global level and thus can be very interesting for various kinds of analysis (Schedl & Hauger, 2012). A user may also publish other relevant information in the tweet, e.g. tags, URL's or links to other users. This can provide a context around their utterance.

Esparza et al. (2010) have researched the Real Time Web as a source of recommendation knowledge in a different way. They focus on the Twitter-like review service Blippr. On this website, users express their views on products like applications, music, movies, books and

games in 160 characters. In the study, an index of movies mentioned on Blippr is made. This index is used to make recommendations based on user profiles and collaborative filtering. A test on the predictive ability of the service shows that 1.35 out of 5 recommended movies are known to be liked by a user. Although this study is not focused on music sharing on micro-blogs, it shows that micro-blog data can be used to make recommendations.

In conclusion, blogs and micro-blogs contain a combination of their own content, tags or keywords to describe that content and links to other content. Music and music-related information can be described by a combination of those aspects.

Improvements of perceived quality

According to Celma (2010), music recommendation algorithms try to make accurate predictions about what a user could listen to. However, the usefulness of a recommendation is not always taken into account. It is proposed that recommender systems should exploit the long tail of popularity in large music collections, i.e. help the user find potentially novel and interesting items. Novelty and relevance are the key elements of useful music recommendation. The measured perceived quality in the study of Celma is neutral or negative for novel recommendations. This shows that a recommender system should provide a rationale for its choices. The study also suggests classifying users into four types of listeners (savant, enthusiast, casual and indifferent). The usefulness of recommendations may vary highly among these types.

Schedl, Hauger & Schnitzer (2012) describe six factors that a new generation of music retrieval systems should take into account:

- Similarity – this is a well-known factor that describes the music similarity in the earlier mentioned content-based, context-based or collaborative way;
- Diversity – although recommended music should be similar to liked music, diversity has to be ensured. Artist or album repetition probably has to be avoided;
- Familiarity/Popularity vs. Hotness/Trendiness – A distinction has to be made between long-term popularity and current trendiness;
- Recentness – This factor describes a certain musical item in terms of closeness to the present;
- Novelty – By this factor, the authors mean novelty in the perspective of the user. A user doesn't want a recommendation system to just present familiar music;
- Serendipity - This addresses the fact that a user should be surprised in a positive way. The effect can occur when the user discovers an unexpected item or an item he/she was not aware of.

A recommendation system that covers these factors can be of high value from a user perspective. The factors can be measured and influenced in several ways. The challenge in this study is to implement these factors using blogs and micro-blogs.

User understanding

Laplante (2010) notices that people, who are searching for music, ascribe particular importance to recommendations from people they know and trust. She proposes inclusion of social networking tools that facilitate the sharing of information between users. In addition, her study shows that music preferences change depending on a user's context (mood, activities, etc.), which emphasizes the need for recommender systems to acknowledge the 'dynamic nature of relevance'. Celma (2010) contributes to this topic of user understanding

by suggesting that a system should understand *why* a user likes a specific kind of music (e.g. does he/she like the pop side or the experimental side of The Beatles?).

A few guidelines can be derived from the above mentioned points of improvement. Recommender systems should provide users with a variety of music, both familiar and novel. Familiarity in terms of long-term popularity may not have the effect of serendipity, while familiarity in terms of trendiness may have that effect. Recommendations that involve novel music need an added context to motivate the recommendation. In this, it might also be useful to create trust between the user and the recommender. Finally, a recommender system should recognize the multiple personalities of a user to optimize the perceived quality.

Combination of information

Passant & Raimond (2008) explain how social music data can be combined with the Semantic Web. Their approach uses the Friend of a Friend approach to link various types of data. Based on a Resource Description Framework (RDF)[iv], which is a standard model for data interchange on the web, a large amount of social networks is linked in order to recommend music-related data. While most of their study is out of the scope of our study since it tries to cover all kinds of music-related data (e.g. cover art and locations of venues), it shows how information across various social networks can be combined by finding links between content. Links can for example result from related people or tags in multiple networks. It can be imagined that an 'author' on Twitter (*a1*) talks about a lot of the same songs as an author of a blog (*a2*). This could mean that a certain user that likes music on the blog of *a2*, might also like music Tweeted by *a1*.

Tsai et al. (2006) describe a model for blog recommendation, using three dimensions:

- Semantic – This covers the content of a blog (post) and tags describing it. Applied to music recommendation we could use the semantic dimension to describe which song and artist are mentioned and even which tags are given to an item;
- Value – This covers user characteristics. In music recommendation, this involves the taste of the user (in both liked music and liked blogs/micro-blogs), but even the context of the user, to take multiple personalities into account;
- Social – This is concerned with all the social links in a blog or article (including comments). Applying this to the current study, it could involve links between blogs and Twitter-users. Social links also occur on Twitter itself, taking into account the follower-followed construction and mentions of other users in Tweets.

This model can be used to link blogs and micro-blogs, which may result in recommendations of content published on an unknown (micro-) blog. The section on music sharing through blogs and micro-blogs already showed similarities in the form of both media-types. These similarities should be leveraged to establish connection between different media.

Evaluation of recommendation systems

Feedback

Celma (2010) distinguishes implicit and explicit feedback in the usage of a recommendation system. Implicit feedback is retrieved from users' listening habits, i.e. the songs a user listens to and play-counts of the songs. Only positive feedback is measured, since the system only keeps track of the tracks that a user did listen to. In contrast, explicit feedback is concerned with ratings, either with a 'like/dislike'-approach or a scaled rating. This measures both positive and negative feedback.

Type of evaluation
Most reviewed literature on music recommendation uses a prediction-based evaluation method for their proposed recommendation approaches. Results of a prototype are compared to results of a trusted recommendation system. Zangerle et al. (2012) compares their computed track recommendations to the recommendations of the *Last.fm* API and calculates the coverage of those recommendations. However, they suggest online user tests as a part of future research to evaluate their approach from a user's point-of-view. Tremblay-Beamont & Aïmeur (2005) describe a music blog recommendation system which is evaluated by comparing predicted ratings to actual user-ratings. This recommendation method goes beyond perceived quality, since it can also measure the accuracy of bad recommendations. A similar evaluation method is used by Esparza et al. (2010) to evaluate movie recommendations based on micro-reviews. However, only liked movies are used to test the prediction of the recommender system.

Celma (2010) does describe an evaluation method to determine perceived quality and usefulness. His system presents a set of songs to the user, using several recommendation approaches. The user is, after giving demographical information, asked whether he/she knows the song already. In the next question, the user is asked for his/her rating on a scale from 1 to 5. Based on these two questions, conclusions can be drawn regarding the different recommendation approaches and the relation between novelty and perceived quality.

A user study is also described in Guy et al. (2010), where a system that recommends social media items is tested. Different recommendation approaches are used and recommendations are presented with or without an explanation. This evaluation method gives a good overview of perceived quality of recommendation approaches and the usefulness of explanations for the recommendations. A combination of the evaluation methods of the studies of Celma and Guy et al. is preferable for the current study. In that way, perceived quality can be measured and the influence of novelty and explanations can provide insights for further improvements.

6. Interviews & Questionnaires

A number of important topics could be selected from the results of the literature review:

- Recommendation approaches (collaborative, content-based, context-based, tags);
- Combination of information from different sources;
- Potential aspects for high quality recommendations (novelty, relevance, mood-analysis, explanations).

These topics were used to formulate questions for experts in the realm of music recommendation.

The insights of the interviewed experts are similar to some suggestions in the literature review and emphasize important aspects of music recommendation. A recommender system should create trust and provide a context to its recommendations. In this way, recommendations become more 'human'. The aspects of the source should be taken into account; blogs can provide qualitative data while Twitter has a bulk of noisy data. Recommendations should not be too similar to what a user already likes.

The results of the user questionnaires also give some suggestions which can be used in the design of a prototype. A recommender system should provide novel recommendations in addition to similar ones, which corresponds to Schedl, Hauger & Schnitzer (2012). It is also important to have a good recommendation source and give an explanation or personal story to complement the recommendation. This is in line with findings of Celma (2010) and the results of the expert questionnaire. No interesting deviations were found in analyzing the different music listening frequencies.

7. Prototype

In this section, the prototype 'Songdice' is described. The prototype was created in three phases: requirements, design and implementation.

Requirements

Several aspects which can establish high quality music recommendation were found in the literature review, the expert opinions and the user opinions. In table 1, an overview presents the mentioned aspects in the different research phases of this study. If the sources in this phase had a negative or more elaborative opinion on the aspect, this is reflected in the table.

	Literature	**Expert Interviews**	**Expert questionnaires**	**User questionnaires**
Hybrid approach	X		X	
Use of tags	X	"Eliminate bad tags"	"Don't use tags"	
Recommend novel items	X		X	X
Provide a rationale	X			X
Classify listeners	X			
Recognize multiple personalities in one user	X		X	
Create trust	X	X		
Combine sources on a semantic, social and value-level	X			
Determine source-quality		X		X
Provide a context			X	X
Make recommendations human		X	X	

Table 1, important aspects of music recommendation, mentioned in the different research phases

A number of aspects were mentioned in multiple phases. Some aspects were considered more important than others. In addition, a number of aspects contain overlap or need a more elaborative explanation. Therefore, the table of important aspects is converted to the following summary of the prototype's requirements:

- Hybrid: multiple recommendation approaches should be combined to overcome barriers of the separate approaches;
- Novelty: recommendations should not be too similar to the music that a user already knows;
- Explanation: the recommender system should rationalize the recommendation of novel items;
- Context: the recommendation should add contextual information. In this way, the item becomes more desirable and a story is added;
- Trust: the recommender should build trust by choosing trustworthy recommendation sources and relating this source to the user. A proposed connection should be close to the user, to increase trust;
- Linkage of data: music-related data should be collected from a number of blogs and micro-blogs. The sources should be linked on a semantic-, social-, and value-based level, following the model of Tsai et al. (2006). The quality of the sources should influence the linkage and thus the recommendations;
- Tags: tags should only be used when they are assigned by experts or represent genres of songs. This eliminates tags without added value.

Besides the requirements for the recommendation functionality of the prototype, a number of requirements can be made regarding the evaluation:

- Ask for explicit feedback to include negative feedback;
- Perceived quality has to be quantified. The evaluation process of Celma (2010), in which users grade the recommendation, is a good example of this;
- Measure the novelty of the recommendation;
- Measure demographic characteristics in order to describe the participants.

Design

The recommendation process

The starting point of the prototype is a difficult aspect, since no user-data is collected yet. Therefore, the prototype will present a set of genres from which the user can select one. The genres are obtained from The Hype Machine, where genre-tags are related to music blogs. A blog is described by multiple genres. The connection between a blog and a genre is either made by people from The Hype Machine or derived from Last.fm. A song from a blog that has posted something in the selected genre will be presented to the user. By this process, not only a starting point is created, but the presented blog can be considered trustworthy. Firan et al. (2007) show that a large number of Last.fm's top tags are genre-related. Since only these (popular) genre tags are used as a starting point, less useful tags (e.g. 'want to see live', 'male vocalist', 'favorite') are eliminated.

The blogs are expected to contain posts on a combination of popular and relatively unknown artists, which adds the aspect of novelty. This novelty of recommendations will be evaluated. Since blogs are described by various genre-tags, users are not stuck to the genre they chose in the beginning. A blog that posts pop, rock and electronic music might present rock music to a user that selected 'electronic' as his favorite genre. The genre of the recommendation is thus not too similar to the user's taste, but is assumed to be related. Since the writer of the blog likes both genres, this approach could add serendipity.

In addition to the song itself, information about the blog and the blog post is given. This adds a context to a recommendation and can increase the trust of a user. The user can rate the quality of the recommendation and tell whether he/she knew the song or the artist already. If the user liked the song (rating is at least 7 out of 10), another song from the same blog will be played. If the user does not like the song, a different blog from the same genre will be selected. The prototype will rationalize its choice, in order to provide an explanation and increase trust.

Each genre is represented by at least three blogs and each blog is represented by at least three songs. If a user has listened to all songs from a particular blog, a new blog in the chosen genre will be selected. If a user rated all songs from the blogs in a particular genre lower than 7, the blog that was given the highest rating by the user will be selected. While the blog is chosen based on the chosen genre and user ratings, a song from that blog is always selected randomly. The only condition is that the user didn't listen to the song yet. Figure 2 shows the flowchart of the selection process of Songdice. In figure 3, a model of the genres, blogs and songs is presented, together with a possible path.

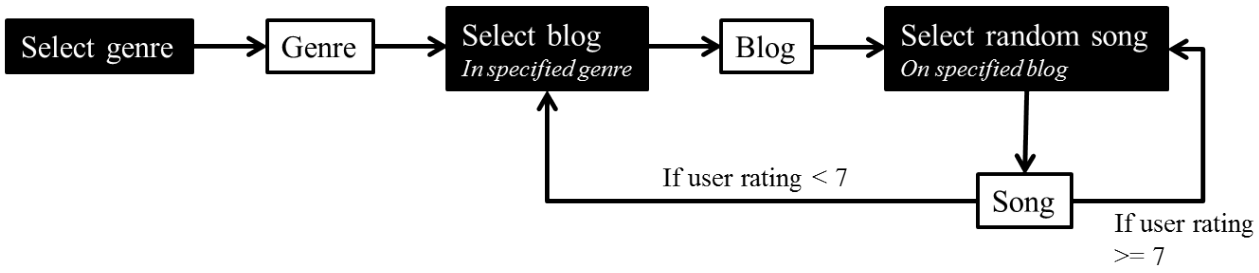
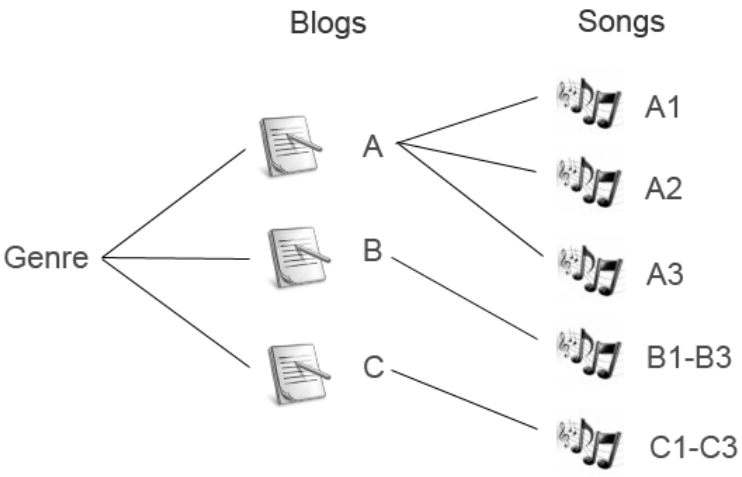


Figure 2, a flowchart of the selection process of Songdice.



Possible path

Genre: rock

A2 (7) - A3 (5) - B2 (4) - C1 (8) - C3 (6) - Electronic B2 (6)

Figure 3, model of the data and a possible recommendation path. The last recommendation is random. The numbers between brackets represent ratings given by the user.

The prototype only uses data from blogs, since they can provide a qualitative context, as several experts have noticed. Twitter is not used in this prototype since it can best be used for the quantity of its data. In the current study, the timeframe was too short to get a well-sized data set with the Twitter Streaming API. Twitter's data is not qualitative enough to add context to a recommendation and is thus not used for that purpose.

After the recommendation of five songs, the user will be given a random recommendation, i.e. in a genre that the user did not choose. This recommendation will be explained by telling the user that the genre is different from what the user indicated to like, but that he/she might like the song. In this way, it seems that the recommender system did think about the recommendation, although it is random. The random song is included to compare the perceived quality of a random song with the perceived quality of a logically recommended song. The user is not expected to like the random song.

The evaluation process

To classify the participants, some demographic data will be obtained prior to the recommendations. In addition to their age and gender, participants are asked how many hours per week they listen to music. They are also asked to determine their musical background, by choosing from options that are used by Celma (2010):

• *None*: no particular interest in music related topics.
• *Basic*: lessons at school, reading music magazines, blogs, etc.
• *Advanced*: regular choir singing, amateur instrument playing, remixing or editing music with the computer, etc.
• *Professional*: professional musician—conductor, composer, high level instrument player—music conservatory student, audio engineer, etc.

After providing this information and choosing their favorite genre from a list of nine genres, the recommendations are presented to the participant. Half of the participants get recommendations with a rationale, the blog post and blog information. This mode will be called the *explanation-mode*. The other half just get the song. This will be called the *no-explanation-mode*. In this way, the effect of context and an explanation in the recommendation is tested. The participants indicate whether they knew the artist or the song before, in order to determine the novelty of the recommendation. A song is rated on a scale from 1 to 10. The effect of a blog-based context and rationale can thus be tested against the novelty and randomness of the recommendation as well as the type of listener.

The user interface of both the *explanation-mode* and the *no-explanation-mode* of Songdice is presented in figure 4.

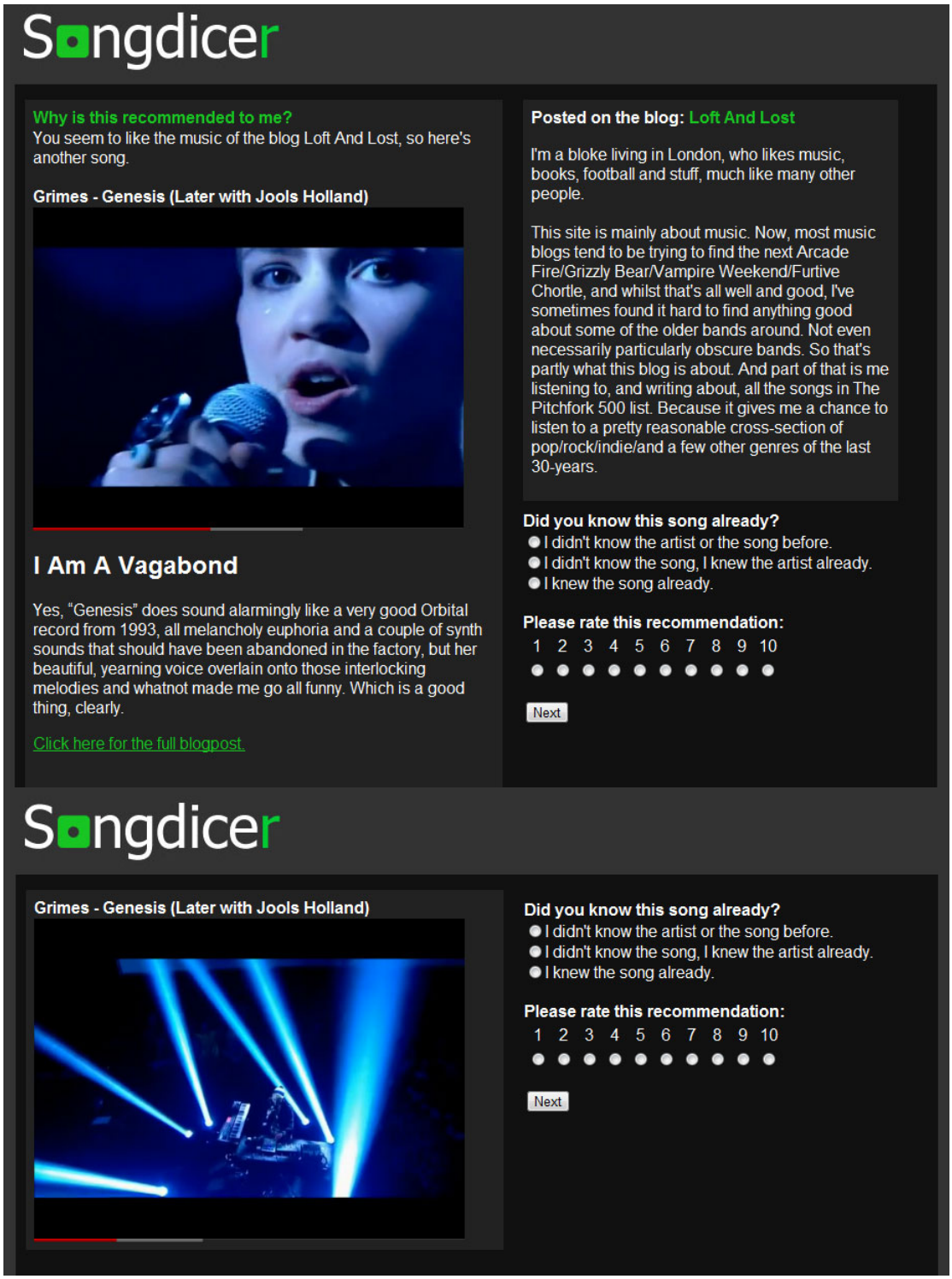


Figure 4, the explanation-mode (top) and the no-explanation-mode (bottom) of Songdice.

Implementation

Prior to developing the system, 69 songs were collected from 17 blogs. The blogs were described by 14 genres, but only nine of those were connected to at least three blogs. These were pop, rock, indie, electronic, house, folk, dance, singer-songwriter and hip-hop. We selected these genres to cover a variety of music tastes and thereby please different users. The blog selection was made by searching for the specific genre on The Hype Machine and analyzing the source of the first blogs that appeared in the list.

For each recommendation, except for the first and sixth, the system reads the last recommendation and decides whether the user wants a new blog. The availability of unrated songs is taken into account by reading all the ratings and the songs from the preferred blog. After selecting a blog, a song is chosen randomly from that blog. When the user rated six songs, the database entry is marked 'finished', in order to only evaluate users that completed the whole process.

8. Hypotheses

The focus of the prototype test in this study is on measuring the influence of blog-based context and rationale in music recommendation. The literature and the results of the questionnaires suggest that these aspects improve the quality of music recommendation. It is expected that the aspects create trust and make recommendation human. According to the literature and the expert interview, this could increase the perceived quality of recommendations too. Perceived quality is regarded as the most important measure since it covers multiple requirements of good music recommendation (e.g. novelty, similarity, serendipity) and it is a user-centered measure. From this line of reasoning, the first and most important hypothesis can be defined:

H1: recommendations in the explanation-mode will have a higher perceived quality than recommendations in the no-explanation-mode.

Since there are different kinds of recommendations and different kinds of users in this test, the hypothesis should be tested in various situations. According to the literature, explanations are especially useful with novel recommendations. In the current study, we can also test the value of explanation for random recommendations. Since these recommendations are presumably perceived as 'wrong' *without* an explanation, but may add serendipity *with* an explanation, we assume that random recommendations have a higher perceived quality when they are explained. In addition to testing H1 on all recommendations, two sub-hypotheses can be developed:

H1a: novel recommendations in the explanation-mode will have a higher perceived quality than novel recommendations in the no-explanation-mode.
H1b: random recommendations in the explanation-mode will have a higher perceived quality than random recommendations in the no-explanation-mode.

The effect of the explanation-mode will also be evaluated for different types of musical backgrounds and different listening frequencies of participants. People with a more advanced musical background might feel more connected to music-bloggers than indifferent people. The difference in quality between explained and not-explained recommendations could thus be bigger.

H2: the improvement of quality through explanation of recommendations is bigger for users with a more advanced musical background.

People who are more into music are probably also more critical towards music. This could decrease the quality of recommendations. This is not contradictive with the previous hypothesis, since it focuses on the absolute rating, rather than the difference that is caused by explanations.

H3: users with a more advanced musical background perceive a lower quality of recommendations than users without that background.

The fourth hypothesis is concerned with novelty of recommendations. Several experts noticed that blog-based recommendations would be based on new trends. In additions, literature defined one role of blogs as 'alternative sources of news and public opinion'. We suggest that most recommendations made by the prototype are novel to the users.

H4: at least half of the number of recommended songs is unknown to the user.

The literature review showed that in general, ratings for novel recommendations are neutral or negative. It can be suggested that familiar songs receive a higher rating and thus a higher perceived quality than novel songs.

H5: novel songs have a lower perceived quality than familiar songs.

Not only hypotheses that try to answer the research question of this thesis can be formulated. The data also allows us to test more general connections. The following hypotheses are based on these connections and are rationalized very concisely:

People who are more interested in music spend more time listening to it.

H6: people with a more advanced music background listen to music more often than people with a less advanced musical background.

Random recommendations may be derived from a blog with a different genre than the user indicated to be his/her favorite genre from the list and can thus be of a genre that the user does not like.

H7: random recommendations have a lower perceived quality than non-random recommendations.

9. Results

93 participants, aged between 16 and 58, completed the test in a time span of two weeks. This resulted in 558 rated songs. 47 participants got recommendations in the explanation-mode, 46 participants got their recommendations in the non-explanation mode. In this section the previously developed hypotheses will be tested, using the results.

Explanation vs. perceived quality

The first hypothesis and its sub-hypotheses are concerned with the effect of rationale and context on the perceived quality of a recommendation. Table 2 shows the mean grade for the explanation mode and the no-explanation mode, resulting from various groupings. The means are found with an Independent Samples T-test. Significance is determined with one-tailed testing, since H_0 ($\mu_1 - \mu_2 = 0$) and H_1 ($\mu_1 - \mu_2 > 0$) where μ_1 is the mean grade in the explanation mode and μ_2 is the mean grade in the no-explanation mode.

	Explanation mode	**No-explanation mode**
All ratings (n=558)	5.60 (2.00)	5.40 (2.13)
Familiar songs (n=42)	5.40 (1.28)	5.18 (1.58)
Novel songs with familiar artists (n=54)	6.07 (1.69)	5.60 (2.31)
Novel songs with unfamiliar artists (n=462)	7.59 (2.00)	7.20 (2.07)
Random recommendations (n=93)	5.28 (2.07)	4.83 (2.01)
Non-random recommendations (n=465)	5.66 (1.99)	5.51 (2.14)
Men (n=324)	5.34 (2.13)	5.39 (2.05)
Women (n=234)	5.92* (1.79)	5.41* (2.24)

*Table 2, mean grades in various groupings, with the standard deviation between brackets. * marks a significant difference ($p < .05$). n is the number of ratings in the group.*

The results show that, although the mean grades in the explanation mode are higher than the mean grades in the no-explanation mode (expect for the group of men), there are no significant differences in the groups of all, novel or random ratings. H1, H1a and H1b are all rejected. However, there is a significant difference in the ratings of women.

Musical background vs. perceived quality

Hypothesis 2 states that the effect of the explanations is bigger when users have a more advanced musical background. Table 3 groups the four types of musical backgrounds and presents the mean grades in the two recommendation modes.

Musical background:	**Explanation mode**	**No-explanation mode**
None (n=144)	5.32 (2.16)	5.05 (2.17)
Basic (n=156)	5.81 (1.99)	5.79 (2.31)
Advanced (n = 234)	5.61 (1.96)	5.52 (1.98)
Professional (n=24)	5.33 (1.63)	4.78 (1.67)

Table 3, mean grades in various groupings, with the standard deviation between brackets. n is the number of ratings in the group.

No significant difference is measured between the two modes, i.e. the explanations had no significant effect. H2 is rejected. In addition, only four participants had a professional musical background, which makes this group too small to make statistical statements about it.

The third hypothesis proposes that users with a more advanced musical background perceive a lower quality of recommendations than users with a less advanced musical background. A one-way ANOVA shows a significant difference ($p < .05$) in the mean grades of the various backgrounds. To further test the linkage of musical background and perceived quality, an Independent Samples T-Test is done between the different backgrounds. The hypothesis is based on an assumption, rather than on the literature review. For this reason, the significance is determined with two-tailed testing. The professional musical background is left out, since it doesn't contain enough participants. Results are shown in table 4.

Musical background:	
None (n=144)	5.16 (2.16)*
Basic (n=156)	5.80 (2.13)*
Advanced (n = 234)	5.57 (1.96)

*Table 4, mean grades of the participants with different musical backgrounds, with the standard deviation between brackets. * marks a significant difference (p < .05). n is the number of ratings in the group.*

The only significant difference can be found between the 'None' and 'Basic' musical backgrounds. The mean grades given by people with a basic musical background are significantly higher than the grades given by people with no musical background. This is opposite to the hypothesis. The 'Advanced' musical background shows no significance with any of the other two backgrounds. H3 is rejected.

Novelty

Of the 558 recommendations, 462 (82.8%) were novel songs to the participants, 54 (9.68%) were songs with familiar artists and 42 recommendations (7.53%) were songs that the participants already knew. Hypothesis 4 is accepted.

The fifth hypothesis states that novel songs have a lower perceived quality than familiar songs. Novel songs have a mean rating of 5.29, songs with a familiar artist have a mean rating of 5.85 and known songs have a mean rating of 7.36. The differences between the three levels of novelty are significant based on a one-tailed Independent Samples T-Test. Songs with a familiar artists are rated higher than novel songs ($p < .05$) and known songs are rated higher than songs with familiar artists ($p < .001$). Based on this, H5 is accepted.

Listening hours

In figure 5, the mean listening hours are shown, grouped by the four different musical backgrounds.

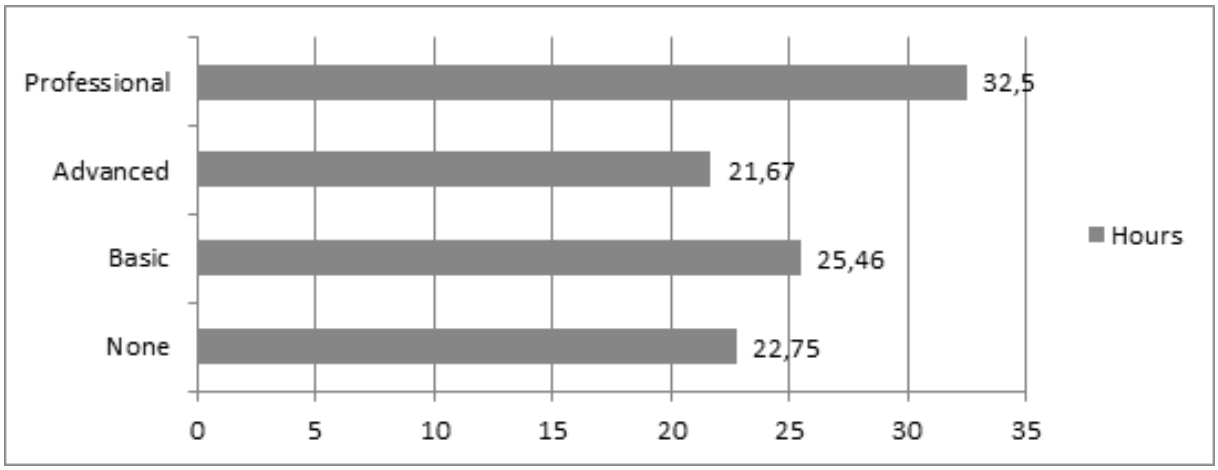


Figure 5, mean listening hours grouped by the musical backgrounds

It can be seen that participants with a more advanced musical background do not necessarily listen to music more often. Again, the group of participants with a professional background was too small to make any statistical statements about. H6 is rejected.

Randomness
The mean rating of random recommendations is 5.59, while the mean rating of the other recommendations is 5.05. This is a significant difference ($p < 0.05$) and thus H7 is accepted.

10. Discussion

Significance
Only three of the hypotheses are accepted, based on the results of the prototype test. This is partly due to the observation that the explanations have no significant effects on the perceived quality of the recommendations. However, the results show higher user ratings in the explanation mode. The insignificance is partly caused by the high standard deviations. This is not solved by leaving out extreme ratings (1 and 10); effects are still insignificant. The low amount of available songs and the very basic algorithm induces a high variance in quality of the songs recommended to the user. A more advanced prototype could filter out songs and blogs after receiving a number of low ratings, which should improve grades and lower standard deviations. In addition, a larger number of participants could make results more reliable and allow for a more diverse analysis of different users.

Validity
The validity of the results can be discussed. The users were asked to rate recommendations, which should reflect their perceived quality. Although the users were informed that these ratings were not just the similarity to the earlier selected genre, it is possible that users still perceived that aspect as the most important criterion. An improved prototype test should use a construct of perceived quality, which contains multiple variables. For example, it might be useful to know why a participant liked or disliked a song. The general perceived quality of the system should also be evaluated. This could reflect the value of added context in a better way. An interesting addition would be the implicit evaluation of recommendations, by measuring for how long a user listens to a particular song.

In measuring the significance of rating differences between the different levels of novelty, an Independent Samples T-Test was used. However, not all samples are independent. Ratings are compared between users, but also within the evaluation of an individual user, since it is possible to that the user knows a selection of the songs. The occurrence and timing of novel and known songs is not determined, which makes it assumable that any dependency effect is eliminated. The effect on the reliability of the study is thus minimal.

A significant effect of the explanations was unexpectedly found among women, while men seem to add no value to the explanations. This gender difference cannot be explained with the reviewed literature and is an interesting topic for further research. This difference could also be caused by a low number of participants.

Evaluation improvements
The classification of the different types of listeners should be improved. In this evaluation, a combination of the musical background and the listening frequency was used. This didn’t lead to findings of significant differences. This is an important point for improvement; when the system creates a better profile of its users, the information can be used to improve both the recommendations and the explanation of the recommendation. A bigger group of participants can lead to better findings, since there were not enough participants with a professional musical background in this study.

Finally, it is worth noticing that a big percentage of the recommendations were novel to the users, but novel recommendations still received a lower rating than familiar songs and artists. This can be due to a minimal added value of the context and rationale complementing the recommendation. However, it can also be possible that a lot of novel recommendations do not have the quality of popular songs. The following improvements section will analyze how this problem could be tackled.

11. Conclusion

We are now ready to answer the sub-questions including the research question. Furthermore, areas for further research can be identified. In previous years, some very interesting developments occurred in the field of music recommendation, involving new approaches and opportunities. While the content-based and collaborative approach can still be seen as the basis of music recommendations, the literature review showed the introduction of the use of social media.

Blogs and micro-blogs are two forms of social media and can be seen as alternative media sources. In this study, experts were asked how they thought that blogs and micro-blogs could improve music recommendation. The results from the interview show that (micro-) blog data potentially adds a context to the published information. When concerned with music, (micro-) blogs are a good source for information in the long tail of artists and songs. However, in recommending content from a certain post, the reliability should be taken into account. Only very influential blogs can serve as a starting point for recommendation and only reliable information will be trusted. (Micro-) blogs can also be used in a quantitative way, where trending music on the web is detected.

Novelty is an important but complex aspect of music recommendation. It can lead to either serendipity or a loss of perceived quality. Trust is an essential factor to achieve a high perceived quality. This statement was supported by literature, potential users and experts. Trustworthy (micro-) blogs should be selected from a user perspective and subsequently trust should be built between the source and the user. This can be done by using the blog's information to add a context and a rationale to recommendations. While we did not quantify trust in this study, it could be quantified by asking users to forward recommendations to their friends or explicitly asking their trust in a source.

In this study, a prototype was built and evaluated by users. Results show that blogs are a good source for novel music and can potentially improve recommendations by adding the context and rationale that was earlier proposed by both experts and potential users. However, the most effective way to use this explanation is still to be found; a significant effect was only found in the group of female participants. Unfortunately, we were not able to explain this difference and this difference should be further researched with more participants. Novel songs were rated significantly lower than familiar songs. A way to improve novel recommendations still needs to be found. The hypothesis that an explanation could increase the perceived quality of a random recommendation was rejected, which emphasizes the fact that further research on effective explanations is needed.

To summarize, a literature review, interviews with potential users and experts and a prototype evaluation showed that blogs and micro-blogs can improve the perceived quality of

recommendations by creating trust, using personalization and exploiting the quality in the long tail.

12. Further Research

From this study, some interesting areas for further research can be identified. The gender difference in the effect of explanations was not mentioned in the reviewed literature and is an interesting topic to study. Different forms of explanations should also be tested in order to determine the most effective form in terms of improving perceived quality. While measuring perceived quality, it would also be useful to quantify the users' trust in sources of recommendations. The perceived quality of and trust in Songdice's recommendations should be compared to collaborative and content-based filtering in order to research the effectiveness of Songdice. Finally, an improved prototype can be evaluated, where the discussion of the current evaluation should be taken into account. This evaluation should ideally involve more participants and data (genres, blogs and songs) than the current study and test the effectiveness of different aspects of the recommendation system.

References

M. Balabanovic & Y. Shoham. Content-Based, Collaborative Recommendation. In *Communications of the ACM*, 1997.

C. Basu, H. Hirsh, & W. Cohen. Recommendation as classification: Using social and content-based information in recommendation. In *Proceedings of the national conference on artificial intelligence*, 1998.

H. Tremblay-Beamont & E. Aïmeur. Jukeblog: A recommender system in music weblogs. In *Proc. of the IADIS Int. Conference on e-Commerce*, 2005.

O. Celma. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer, 2010.

S. Esparza, M. O'Mahony, B. Smyth. On the real-time web as a source of recommendation knowledge. In *Proc. of the 4th ACM Conference on Recommender Systems*, 2010.

C. Firan, W. Nejdl, R. Paiu. The Benefit of Using Tag-Based Profiles. In *Web Conference*, 2007.

I. Guy, N. Zwerdling, I. Ronen, D. Carmel, E. Uziel. Social Media Recommendation based on People and Tags. In *Proc. SIGIR*, 2010.

S. Herring, L. Scheidt, S. Bonus, L. Wright. Bridging the Gap: A Genre Analysis of Weblogs. In *Proceedings 37th Annual HICSS Conference*, 2004.

A. Laplante. Users' relevance criteria in music retrieval in everyday life: An exploratory study. In *Proc. ISMIR*, 2010.

X. Li, L. Guo, Y. Zhao. Tag-based Social Interest Discovery. In *Proc. WWW*, 2008.

A. Passant & Y. Raimond. Combining Social Music and Semantic Web for music-related recommender systems. In *Social Data on the Web Workshop*, 2008

M. Schedl. On the Use of Microblogging Posts for Similarity Estimation and Artist Labeling. In *Proc. ISMIR*, 2010.

M. Schedl & D. Hauger. Mining Microblogs to Infer Music Artist Similarity and Cultural Listening Patterns. In *Proc. WWW*, 2012.

M. Schedl, D. Hauger, D. Schnitzer. A Model for Serendipitous Music Retrieval. In *2nd Workshop on CaRR*, 2012.

T. Tsai, C. Shih & S. Chou. Personalized blog recommendation using the value, semantic, and social model. In *Innovations in Information Technology*, 2006.

E. Zangerle, W. Gassler, G. Specht. Exploiting Twitter's Collective Knowledge for Music Recommendations. In *Making Sense of Microposts*, 2012.

Endnotes

[i] www.shuffler.fm
[ii] www.hypem.com
[iii] www.alexa.com/topsites, retrieved in May, 2012
[iv] www.w3.org/RDF/