Exploiting Surface Features for the Prediction of Podcast Preference

Manos Tsagkias[1], Martha Larson[2], and Maarten de Rijke[1]

[1] ISLA, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
e.tsagkias@uva.nl, mdr@science.uva.nl
[2] Information and Communication Theory Group, Faculty of EEMCS, Delft University of Technology, The Netherlands
m.a.larson@tudelft.nl

Abstract. Podcasts display an unevenness characteristic of domains dominated by user generated content, resulting in potentially radical variation of the user preference they enjoy. We report on work that uses easily extractable surface features of podcasts in order to achieve solid performance on two podcast preference prediction tasks: classification of preferred vs. non-preferred podcasts and ranking podcasts by level of preference. We identify features with good discriminative potential by carrying out manual data analysis, resulting in a refinement of the indicators of an existent podcast preference framework. Our preference prediction is useful for topic-independent ranking of podcasts, and can be used to support download suggestion or collection browsing.

1 Introduction

A podcast is an audio series made available on the internet via subscription [4]. Podcasts are not always a product of professional producers. Rather, they can be published by individual users or by companies and institutions such as government bodies or museums. Diverse origins, content and production methods characterize the podosphere, the totality of podcasts available on the internet. Associated with this variation is the phenomenon that different podcasts enjoy different levels of appeal among listeners. Independently of topic, certain podcasts are preferred by users above other podcasts. In our work, we address the task of predicting podcast preference, in particular, of classifying podcasts as preferred or non-preferred and of ranking podcasts by preference.

Conventionally, users access podcasts by subscribing to a feed using a podcast aggregator [4]. Episodes are then automatically downloaded as they are published and stored on a portable audio player for later listening. It is also possible to listen to a podcast while sitting at the computer or to selectively download particular episodes instead of subscribing to the podcast feed. What is shared by these listening scenarios is that as a first step in the process listeners must identify a podcast that they are interested in listening to. There are multiple routes by which a listener gets matched up with a podcast [2]. Common scenarios are that the podcast is either suggested by a person or a website,

M. Boughanem et al. (Eds.): ECIR 2009, LNCS 5478, pp. 473–484, 2009.

discovered via browsing or found using a search engine. Our podcast preference prediction framework is applicable in all of these scenarios. Projected preference will support mining the podosphere to discover new podcasts with high potential to become popular. These podcasts can be "featured" in engines and portals, as is already common practice. Additionally, projected preference can help discriminate between podcasts that have not drawn user attention because they are inherently unappealing and podcasts that simply have yet to be discovered.

A source of information that can potentially be used for preference prediction are user ratings, either explicit or implicit (e.g., through download statistics). However, such information is not always available, for instance, for reasons of privacy, confidentiality, or business competition or because a podcast might be too new or its content too obscure to have generated a reliable amount of user ratings. Our goal is to be able to predict preference without relying on user ratings, so that our prediction methods can, e.g., be used during the infancy of a podcast in the podosphere.

The features that we do consider in this paper are surface features, i.e., properties of documents that are observable at the surface, and do not encode information about document content or meaning [3]. Examples of surface features that we consider include length and regularity, and, as will be discussed in Section 4, details related to technical execution and concerning how a podcast is packaged and distributed. These features are derived from characteristics of podcast preference that were identified in a human analysis previously conducted by the authors of [9], whose main findings we corroborate and refine in Section 3.

This work makes several contributions. First, it introduces the problem of predicting podcast preference. Second, it provides a set of surface features, based on a human analysis of characteristics of popular and unpopular podcasts, that can be used to predict podcast preference. Third, it provides an evaluation of preference prediction over 250 podcasts, comparing a set of 5 classifiers. For our experimental evaluation, we narrow our domain from the podosphere at large to the portion of the podosphere listed in iTunes.[1] This restriction allows us to make use of the iTunes "popularity bars" as ground truth for the quantitative evaluation of our approach. Fourth, it provides an analysis indicating which features and methods are most effective for predicting podcast preference.

In the next section, we discuss work related to our research. Then, we revisit and refine [9]'s human analysis of podcast preference, on top of which we formulate surface features for predicting podcast preference in Section 4. We go on to describe the data set and to report on experiments on classification and ranking. Finally, we present a discussion of the results, including an analysis of podcasts for which our approach fails, and an outlook on future work.

2 Related Work

The theoretical foundation of our work is the vast literature on issues of credibility and quality of media, especially of the literature on non-traditional media such as internet content, overviewed in [7]. Issues of credibility and appeal in the

[1] `http://www.apple.com/itunes/`

blogosphere involve user perceptions of the reliability of primary source information embedded in a social network [8, 10] and we consider many aspects of blog preference to have relevance in our research. Our work builds on the *PodCred* framework for assessing the credibility and quality of podcasts presented in [9]. PodCred established a list of indicators for podcast preference divided into the categories *Podcast Content*, *Podcaster*, *Podcast Context* and *Technical Execution*, but stopped short of encoding these indicators into features and exploiting them for automatic preference assessment. Here we build on the PodCred framework, concentrating especially on *Technical Execution* indicators that are readily extractable; see Section 3 for further details on the PodCred framework.

Our work on automatic determination of podcast preference is related to research in the area of text-based (user generated) content. In the domain of user-supplied reviews, automatic assessment of how helpful reviews are to users has been carried out using structural, lexical, syntactic, semantic and metadata features [5]. In the domain of on-line discussions, the quality of posts has been automatically assessed using a combination of features from categories designated: surface, lexical, syntactic, forum specific and similarity [12]. Community-based answers to questions have also been automatically assessed for quality, expressed as *user satisfaction* [1, 6]. Other related work includes research which has investigated the exploitation of topic independent information for improving the quality of information retrieval. In the domain of blogs, features encoding post-level and blog-level credibility indicators have been used as (query-independent) priors to help improve blog post retrieval effectiveness [11]. In particular, the work reported here seeks to exploit the contributions of surface features of podcasts to the problem of predicting podcast preference. In the domain of multimedia, surface features such as length and temporal patterns have been shown to contain useful information for retrieval [13].

3 Characteristics Indicative of Podcast Preference

The PodCred analysis framework presented in [9] comprises a list of descriptive indicators of user-preferred podcasts. We adopt PodCred's indicators as a basis for the features to be used for automatic podcast preference prediction. The PodCred framework was based on a study of user-preferred podcasts only. For our work, we are interested in identifying indicators that have potential to discriminate preferred and non-preferred podcasts. For this reason, we revisit the PodCred framework instead of adopting its indicators off the shelf. In order to confirm and refine the PodCred framework, we carry out a human analysis of non-preferred podcasts. As our data set we choose 16 podcasts that land at the bottom of the list when podcasts in iTunes are ranked by bar-count in the column headed "Popular." We consider each of the podcasts in turn, looking at the feeds and listening to selected episodes, and recording the presence of indicators in each of the four categories, *Podcast Content*, *Podcaster*, *Podcast Context* and *Technical Execution*, of the PodCred framework.

In Table 1 in the column labeled "Non-Preferred," we report the percentage of podcasts found to display each indicator. The statistics reported in [9] are included in the column labeled "Preferred." Preferred and non-preferred podcasts

Table 1. Percentage of non-preferred and preferred podcasts displaying indicators proposed in [9]. The percentages in the third column are taken from [9].

Observed indicator	**% of non-preferred podcasts**	**% of preferred podcasts**
Category Podcast Content		
Topic podcasts	44	68
Topic guests	25	42
Opinions	50	74
Cite sources	19	79
One topic per episode	56	47
Consistency of episode structure	25	74
Interepisode references	0	42
Category Podcaster		
Fluent	25	89
Presence of hesitations	44	37
Normal speech speed	44	42
Fast speech speed	0	53
Slow speech speed	19	5
Clear diction	50	74
Invective	13	5
Multiple emotions	0	21
Personal experiences	56	79
Credentials	25	53
Affiliation	56	21
Podcaster eponymous	13	53
Category Podcast Context		
Podcaster addresses listeners	6	79
Episodes receive many comments	0	79
Podcaster responds to comments	6	47
Links in metadata/podcast portal	13	68
Advertisements	13	53
Forum	6	53
Category Technical Execution		
Opening jingle	31	84
Background music	25	37
Sound effects	25	42
Editing effects	31	53
Studio quality recording	31	68
Background noise	31	26
Feed-level metadata	75	95
Episode-level metadata	50	84
High quality audio	38	68
Feed has a logo	13	58
Associated images	19	58
Simple domain name	38	74
Podcast portal	63	84
Logo links to podcast portal	0	37

can be seen to be characterized by quite distinct trends regarding the indicators that they display. The comparison suggests that the PodCred indicators will be useful as the basis for automatic podcast preference prediction. Particularly striking characteristics of non-preferred podcasts uncovered by the human analysis were their low audio quality, lack of evidence of interaction between podcaster and listeners, and lack of an adequate platform for such interaction (i.e., no commenting facilities or forum). The analysis led to the discovery that podcast episode length tends to be short for non-preferred podcasts. One important example is that of cases of a feed being used to deliver a set of audio files that were created not as a series, but rather for diverse purposes, e.g., a collection of otherwise unrelated recordings by children in a school class.

4 Features for Predicting Podcast Preference

For our approach to prediction of podcast preference, we select indicators from the PodCred framework to transform into extractable features useful for further experimentation with classification and ranking. We focus on indicators that are easily extracted from feeds and represent surface characteristics of podcasts. We choose four indicators from the category *Technical Execution* of the PodCred framework: *Feed-level metadata*, *Episode-level metadata*, *Feed has a logo* and *Logo links to podcast portal*. In the results of the human analysis of podcasts reported in Table 1, these are four of the indicators displaying a radical contrast of occurrence distribution between non-preferred and preferred podcasts. This contrast suggests that these indicators make good features for classification and ranking. We exhaust this potential and leave exploration of the use of less promising features that are challenging or require relatively more computational capacity to extract (i.e., one topic per episode or presence of hesitations in the speech of the podcaster) to future work. Additionally, we include in our selected set the indicator "Regularity," which reflects the temporal publication pattern of a podcast. Regularity is an indicator in the *Podcast Content* category of the PodCred framework, but is not included in the human analysis, which was carried out entirely by hand and for this reason did not include counting publication dates or intervals along the feed lifetime. Finally, we used the indicator "Podcast episode length," which emerged in the human analysis as potentially well correlated with whether or not a podcast is preferred.

In Table 2, the selected indicators are listed, each followed by the specific features that were chosen to encode them. Each feature is listed with its name, a short description and its type. Features are divided into groups depending on the level at which they describe the podcast. Features encoding properties of the podcast as a whole are marked with the level *Feed*. Features encoding properties of the individual podcast feed items are marked with the level *Episode*. Finally, features encoding properties of the feed enclosure, the actual podcast episode audio file, are marked with level *Enclosure*. Grouping the features in this way allows us to design classification and ranking experiments that focus on features derived by considering the podcast as a whole, or, alternatively, samplings of its component parts. Next, we briefly describe the motivation for choices made when we established the indicator to feature mapping in Table 2.

Table 2. Mapping of indicators selected for further experimentation onto extractable features. Features are grouped into levels, according to whether they encode properties of the podcast as a whole (Feed) or of its parts (Episode, Enclosure).

Feature	Level	Description	Type
Indicator: *Feed-level metadata*			
feed_has_description	Feed	Feed has a description	Nominal
feed_descr_length	Feed	Feed description length in characters	Integer
feed_authors_count	Feed	Number of unique authors in feed	Integer
feed_has_copyright	Feed	Feed is published under copyright	Nominal
feed_categories_count	Feed	Number of categories listing the feed	Integer
feed_keywords_count	Feed	Number of unique keywords used to describe the feed	Integer
Indicator: *Episode-level metadata*			
episode_authors_count	Episode	Number of unique authors in episode	Integer
episode_descr_ratio	Episode	Proportion of feed episodes with description	Real
episode_avg_descr_length	Episode	Avg. length of episode description in feed	Real
episode_title_has_link2page	Episode	Number of episodes with titles linking to an episode page	Integer
Indicator: *Feed has a logo*			
feed_has_logo	Feed	Feed has an associated image logo	Nominal
Indicator: *Logo links to podcast portal*			
feed_logo_linkback	Feed	Feed logo links back to podcast portal	Nominal
Indicator: *Regularity*			
feed_periodicity	Feed	Feed period in days	Real
feed_period_less1week	Feed	Feed has a period less than 1 week	Nominal
episode_count	Episode	Number of episodes in the feed	Integer
enclosure_count	Enclosure	Number of enclosures in the feed	Nominal
more_2_enclosures	Enclosure	Feed contains >2 enclosures	Nominal
enclosure_past_2month	Enclosure	Was an episode released in past 60 days?	Integer
Indicator: *Podcast episode length*			
enclosure_duration_avg	Enclosure	Avg. episode duration in seconds (reported in feed)	Real
enclosure_filesize_avg	Enclosure	Avg. enclosure file size in bytes (reported in feed)	Real

Indicators involving metadata reflect the amount of care that is invested into the production of a podcast. Feed-level metadata remains relatively static over time and are likely to have high utility for preference prediction since feed-metadata related features can be extracted without protracted monitoring of the feed. We capture not only presence but also length of the description as well as the effort invested in multi-author collaboration and in associating the feed with keywords and categories that will allow it to be more easily found. Episode-level metadata again reflects podcaster care, with the additional requirement that the effort must be sustained as the podcast continues to be published. Features related to the podcast logo are straightforward to extract and reflect not only

care, but also the intent to build a listenership and establish a community. We make the following choices when translating the regularity indicator into features useful for classification and ranking. In order to discover the feed periodicity, i.e., the length of the release cycle of a podcast, the episode release dates are considered as a time series with start date being the date of the most recent episode and end date 6 months before it. If the feed does not span 6 months, the end date is set to be the date of the oldest episode. For feeds with less than 3 episodes, the periodicity was not calculated but was assigned an arbitrary large number (183 days). From the Fast Fourier Transform on the feed time series we take the weighted average of the five strongest coefficients and extract the resulting period (`feed_periodicity`). In order to be able to determine the effectiveness of different feature choices, we include less complex encodings of regularity among our features. E.g., we include a feature that requires the release period to be less than two weeks, as well as features that reflect recency and raw counts of releases. Last, we include two features that are variants on an encoding of podcast episode length.

5 Experimental Setup

In addressing the podcast preference prediction problem, we concentrate on developing features and combinations of features that can be used for preference prediction and not on developing or optimizing machine learning techniques. In this respect, our goals are comparable to those of [1, 6]. In particular, we want to know the effectiveness of our complete set of features, of individual features, and of features grouped by level (feed, episode, enclosure), both for classifying podcasts as *Popular* or *Non-Popular* and for ranking podcasts.

To answer these research questions, we conduct both classification (Section 6) and ranking (Section 7) experiments. The data set used consists of a set of 250 podcasts feeds comprising of 9,128 episodes with 9,185 enclosures, adding up to ∼2,760 hours of audio. We chose these feeds from a snapshot dated late August 2008 of the feeds listed in each of the 16 topical categories of iTunes (see footnote 3). For each category, we took the feeds in the order they are listed when they are sorted in iTunes using the column labeled "Popular." We then gathered the ten feeds at the top of the list and the ten feeds at the bottom list using a crawler implemented based on the SimplePie[2] library, which allows for RSS parsing. Feeds in non-Western languages, feeds containing video enclosures and feeds that were unreachable were discarded.

For our experiments, we make use of the Weka toolkit [14], choosing to compare a Naive Bayes classifier, with an SVM classifier and several decision tree classifiers — a set representative of the state-of-the-art in classification. All classification results reported were calculated using ten-fold cross validation.

Ground truth was established as follows. We take the ranking yielded by sorting on the iTunes "Popular" column to be indicative of user preference and use this ranking as the ground truth in our experiments. Although the exact mechanism by which iTunes calculates "Popular" is not public knowledge, we make the

[2] `http://simplepie.org`

assumption that it is related to the number of downloads, and, as such, reflects user preference for certain podcasts. For our classification experiments, we build two sets *Popular* and *Non-Popular* by taking the top ten and the bottom ten entries from the Popular-sorted iTunes list for each of the 16 categories. Of the 250 podcasts yielded by our podcast crawl 148 are iTunes-Popular podcasts and 102 iTunes-Non-Popular.

6 Classification Experiments

The first podcast preference prediction experiment we carry out undertakes a binary classification of podcasts into the classes *Popular* and *Non-Popular*. Our initial set of classification experiments explores the individual contribution of each feature listed in Table 2. In Table 3, classification results are reported for runs using a single feature. A classifier that assigns all podcasts to the most frequent class (Popular) achieves an F1 score of 0.74 and is used as a random

Table 3. F1 scores for Naive Bayes, Support Vector Machine (SMO), and tree classifiers (J48, RandomForest, and RandomTree) using a single feature. Boldface indicates improvement over random baseline.

Feature	**F1**				
	Naive-Bayes	**SVM**	**J48**	**Random-Forest**	**Random-Tree**
Random Baseline	0.74				
Level: *Feed*					
feed_has_logo	**0.78**	**0.78**	**0.78**	**0.78**	**0.78**
feed_logo_linkback	0.70	0.72	0.74	0.71	0.71
feed_has_description	0.74	0.73	0.73	0.74	0.74
feed_descr_length	0.50	0.72	**0.76**	0.63	0.66
feed_categories_count	0.38	0.74	**0.78**	0.74	0.74
feed_keywords_count	0.30	0.74	**0.77**	0.68	0.70
feed_has_copyright	0.73	0.73	0.73	0.73	0.73
feed_authors_count	0.74	0.74	**0.77**	**0.77**	**0.76**
feed_periodicity	**0.75**	**0.75**	0.68	0.66	0.66
feed_period_less1week	0.71	0.71	0.71	0.71	0.71
Level: *Episode*					
episode_descr_ratio	0.74	0.73	0.74	0.74	0.74
episode_avg_descr_length	0.38	0.74	0.73	0.60	0.60
episode_title_has_link2page	0.32	0.74	0.73	**0.77**	**0.76**
episode_count	0.46	0.74	**0.79**	**0.76**	**0.75**
episode_authors_count	**0.78**	**0.79**	**0.79**	**0.79**	**0.79**
Level: *Enclosure*					
enclosure_count	0.45	0.74	**0.78**	**0.77**	**0.78**
more_2_enclosures	**0.76**	**0.76**	**0.76**	**0.76**	**0.76**
enclosure_past_2month	0.69	0.69	0.67	0.67	0.69
enclosure_duration_avg	0.57	0.74	0.74	0.71	0.71
enclosure_filesize_avg	0.73	0.74	0.74	0.60	0.61

baseline. In general, tree-based methods out-performed NaiveBayes and SVM, with RandomForest yielding the best performance by a slim margin. Of the 20 features we test, half fail to achieve classification performance above that of the random baseline when used individually. However, among the half that do achieve improvements, there are a several strong performers that show improvements for all classifiers, namely, `feed_has_logo`, `episode_authors_count` and `more_2_enclosures`.

Our further classification experiments investigate which features are potentially most damaging to classification performance. In Table 4, classification results are reported for runs using all features but one, testing omission of each feature in turn. Boldface indicates those cases in which removal of an individual feature improves performance over using all features. No single feature emerges as being particularly detrimental. In other words, in no case does removing a feature lead to performance improvement across the board. RandomForest is generally the best performing classifier and achieves a peak F1 score of 0.83 when using all features except `feed_authors_count`.

Table 4. F1 score for Naive Bayes, Support Vector Machine (SMO), and tree classifiers (J48, RandomForest, and RandomTree) omitting a single feature. Boldface indicates improvement in performance for the respective classifier compared to all features, all levels.

Feature omitted	**F1**				
	Naive-Bayes	**SVM**	**J48**	**Random-Forest**	**Random-Tree**
None - All features, all levels	0.54	0.79	0.76	0.83	0.76
Level: *Feed*					
feed_has_logo	0.53	0.76	**0.77**	0.81	0.72
feed_logo_linkback	0.54	0.79	0.75	0.82	0.71
feed_has_description	0.54	0.79	0.76	0.81	0.71
feed_descr_length	0.53	0.79	0.76	0.82	0.71
feed_categories_count	**0.56**	0.77	**0.80**	0.77	0.75
feed_keywords_count	0.54	0.77	**0.81**	0.82	0.71
feed_has_copyright	**0.55**	0.78	**0.78**	0.81	0.76
feed_authors_count	0.54	0.78	0.76	0.83	0.72
feed_periodicity	0.53	0.78	**0.77**	0.83	0.76
feed_period_less1week	0.53	0.76	**0.78**	0.78	0.71
Level: *Episode*					
episode_descr_ratio	0.54	0.78	**0.77**	0.81	**0.77**
episode_avg_descr_length	**0.55**	0.78	**0.77**	0.81	0.73
episode_title_has_link2page	**0.58**	0.78	**0.80**	0.80	0.74
episode_count	**0.59**	0.78	0.76	0.81	0.74
episode_authors_count	0.54	0.76	**0.77**	0.80	0.72
Level: *Enclosure*					
enclosure_count	**0.59**	0.78	**0.77**	0.81	0.73
more_2_enclosures	0.54	0.79	0.76	0.82	0.71
enclosure_past_2month	0.54	0.79	0.74	0.81	0.71
enclosure_duration_avg	0.53	0.78	**0.77**	0.81	0.71
enclosure_filesize_avg	0.54	0.77	0.76	0.81	0.72

Table 5. F1 score for Naive Bayes, Support Vector Machine (SMO), and tree classifiers (J48, RandomForest, and RandomTree) using a single group of features and all features

	F1				
	Naive-Bayes	**SVM**	**J48**	**Random-Forest**	**Random-Tree**
All Feed features	0.53	0.76	0.69	0.75	0.72
All Episode features	0.44	0.79	0.81	0.79	0.74
All Enclosure features	0.73	0.69	0.79	0.78	0.75
All features, all levels	0.54	0.79	0.76	**0.83**	0.76

Our final classification experiments investigate the contributions of features describing the podcast at various levels. Table 5 reports results of experiments where classification is performed using all features of a single level as well as all features taken together. No particular feature level grouping rivals the use of all features from all levels, although enclosure-level features do show good performance across the board. The top F1 score of 0.83 is achieved by the RandomForest classifier when all features from all levels are used together. NaiveBayes performs relatively poorly, quite possibly a reflection of dependencies between the features used – especially likely for features derived from the same preference indicator, which are potentially rather highly correlated. The RandomForest classifier consistently displays the best performance as it seems to be able to isolate helpful features from our feature set.

7 Ranking Experiments

The second podcast preference prediction experiment involves ranking the top ten podcasts in each of the 16 iTunes topic categories.[3] The goal is to rank the podcasts in order of their "Popular" ranking in iTunes. For ranking purposes we use the RandomForest classifier and all features, which produced the best run (F1 0.83) in the classification experiments. Investigation of the iTunes "Popular" ranking revealed that the very highest podcasts are displayed with a considerable number of popularity bars, and, that for podcasts below rank 3 this number quickly trails off. If we want to emulate the iTunes ranking, our goal should be to produce ranked lists that land iTunes-Popular podcasts in top positions. For this reason, we evaluate the results of our ranking experiment (Table 6) in terms of Mean Reciprocal Rank (MRR), Precision at 3 (P@3) and Precision at 5 (P@5) averaged across all 16 categories.

Note that our ranking algorithm succeeds in landing top iTunes-Popular podcasts at top ranks even though it does not faithfully reproduce the entire ranking, as reflected by the fact that neither Pearson's correlation ρ nor Kendall's τ revealed significant correlation between our top ten ranked lists and those of iTunes (with values of -0.0277 and 0.0227, respectively).

[3] The 16 topical categories in iTunes are *TV and Film*, *Technology*, *Sports and Recreation*, *Society and Culture*, *Science and Medicine*, *Religion*, *News and Politics*, *Music*, *Kids and Family*, *Health*, *Government and Organisations*, *Games and Hobby*, *Education*, *Comedy*, *Business*, and *Arts*.

Table 6. Mean Reciprocal Rank (MRR), Precision at 3 (P@3) and Precision at 5 (P@5) averaged across all categories

	MRR	**P@3**	**P@5**
mean	0.49	0.23	0.51
median	0.33	0.33	0.60
min	0.14	0.00	0.20
max	1.00	0.67	0.80

8 Conclusion and Outlook

We have shown that podcast preference can be predicted by making use of easily extractable features reflecting surface properties of podcasts, especially of features involving metadata completeness and consistency and care of technical execution. The features used for classification were chosen from a set of preference indicators that was adopted from previous work and then extended by further human analysis of podcasts in order to ensure its suitability for the task of differentiating preferred from non-preferred podcasts. We report results from both classification and ranking experiments performed on a group of podcasts listed by iTunes. We are able to separate iTunes Popular podcasts from Non-Popular ones and also rank podcasts such that leading Popular podcasts on iTunes land at the top of the list.

In order to better understand our experimental results, we perform a failure analysis on those podcasts misclassified by our classifiers. The set of iTunes-Popular podcasts includes podcasts that only keep the most current item on the feed and store older items in an archive. Such podcasts tend to be misclassified as Non-Popular, quite likely because in these cases we cannot reliably calculate features related to release regularity. Also, iTunes-Popular podcasts include examples of podcasts no longer currently publishing, but whose topic is timeless (e.g., knitting) so that they don't go out of date. Again, our method tends to classify these as Non-Popular, probably because they lack a recent release. Our larger goal is to extend our approach to encompass indicators from the *Podcast Content*, *Podcaster* and *Podcast Context* categories of the PodCred framework. We expect that given the solid performance of surface features reported in this paper, it will be a challenge to find additional features that yield improvement. Future work will also involve optimizing feature encodings and performing more detailed search for top performing feature combinations.

As we continue to develop methods for predicting podcast preference, we will start to look to applications such as podcast recommendation or collection browsing support. In particular, we are interested in applications in which something is known of the user profile. During failure analysis, we noticed that many false positives seemed quite appealing and displayed a full range of preference indicators from the PodCred framework. These cases were often podcasts of interest to a certain locality, e.g., targeted at residents of a particular city. They also included podcasts published in non-English languages. A readily available explanation for this behavior is that our classifier is identifying podcasts that would be preferred within certain communities, but, because they are not mainstream, do not achieve

the broad exposure necessary to accrue Popular status in iTunes. In the long term, we believe that our methods hold promise to support the exposure and findability of community-targeted and nascent podcasts, providing listeners with a wider variety of preferred podcasts.

Acknowledgments. This research was supported by the E.U. IST program of the 6th FP for RTD under project MultiMATCH contract IST-033104, and by the Netherlands Organization for Scientific Research (NWO) under project numbers 220-80-001, 017.001.190, 640.001.501, 640.002.501, 612.066.512, STE-07-012, 612.061.814, 612.061.815.

Bibliography

[1] Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media, with an application to community-based question answering. In: Web Search and Data Mining (WSDM), pp. 183–194 (2008)
[2] Besser, J.: Incorporating User Search Goal Analysis in Podcast Retrieval Optimization. Master's thesis, Saarland University (2008)
[3] de Jong, F.M.G., Westerveld, T., de Vries, A.P.: Multimedia search without visual analysis: The value of linguistic and contextual information. IEEE Transactions on Circuits and Systems for Video Technology 17(3), 365–371 (2007)
[4] Geoghegan, M., Klass, D.: Podcast solutions: The complete guide to podcasting. In: Friends of ED (2005)
[5] Kim, S.-M., Pantel, P., Chklovski, T., Pennacchiotti, M.: Automatically assessing review helpfulness. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 423–430 (2006)
[6] Liu, Y., Bian, J., Agichtein, E.: Predicting information seeker satisfaction in community question answering. In: SIGIR 2008: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 483–490 (2008)
[7] Metzger, M.J., Flanagin, A.J., Eyal, K., Lemus, D.R., McCann, R.: Credibility in the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment, pp. 293–335. Lawrence Erlbaum, Mahwah (2003)
[8] Mishne, G.: Applied Text Analytics for Blogs. PhD thesis, University of Amsterdam (2007)
[9] Tsagkias, M., Larson, M., Weerkamp, W., de Rijke, M.: Podcred: A framework for analyzing podcast preference. In: Second Workshop on Information Credibility on the Web (WICOW 2008), Napa Valley. ACM, New York (2008)
[10] van House, N.: Weblogs: Credibility and collaboration in an online world (2002) (unpublished ms.)
[11] Weerkamp, W., de Rijke, M.: Credibility improves topical blog post retrieval. In: HLT-NAACL, pp. 923–931 (2008)
[12] Weimer, M., Gurevych, I., Mühlhüser, M.: Automatically assessing the post quality in online discussions on software. In: ACL 2007 Demo and Poster Sessions, pp. 125–128 (2007)
[13] Westerveld, T., de Vries, A., Ramírez, G.: Surface features in video retrieval. In: Adaptive Multimedia Retrieval: User, Context, and Feedback, pp. 180–190. Springer, Heidelberg (2006)
[14] Witten, I.H., Frank, E.: Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers Inc., San Francisco (2005)