Predicting Podcast Preference: An Analysis Framework and its Application

Manos Tsagkias
ISLA, University of Amsterdam, Science Park 107, Amsterdam, The Netherlands. E-mail: e.tsagkias@uva.nl

Martha Larson
EEMCS, Delft University of Technology, Delft, The Netherlands. E-mail: m.a.larson@tudelft.nl

Maarten de Rijke
ISLA, University of Amsterdam, Science Park 107, Amsterdam, The Netherlands. E-mail: mdr@science.uva.nl

Finding worthwhile podcasts can be difficult for listeners since podcasts are published in large numbers and vary widely with respect to quality and repute. Independently of their informational content, certain podcasts provide satisfying listening material while other podcasts have little or no appeal. In this paper we present PodCred, a framework for analyzing listener appeal, and we demonstrate its application to the task of automatically predicting the listening preferences of users. First, we describe the PodCred framework, which consists of an inventory of factors contributing to user perceptions of the credibility and quality of podcasts. The framework is designed to support automatic prediction of whether or not a particular podcast will enjoy listener preference. It consists of four categories of indicators related to the *Podcast Content*, the *Podcaster*, the *Podcast Context*, and the *Technical Execution* of the podcast. Three studies contributed to the development of the PodCred framework: a review of the literature on credibility for other media, a survey of prescriptive guidelines for podcasting, and a detailed data analysis. Next, we report on a validation exercise in which the PodCred framework is applied to a real-world podcast preference prediction task. Our validation focuses on select framework indicators that show promise of being both discriminative and readily accessible. We translate these indicators into a set of easily extractable "surface" features and use them to implement a basic classification system. The experiments carried out to evaluate system use popularity levels in iTunes as ground truth and demonstrate that simple surface features derived from the PodCred framework are indeed useful for classifying podcasts.

Introduction

Podcasts are audio series published online. As new episodes of a podcast are created, they are added to the podcast feed and are distributed over the Internet (Patterson, 2006; van Gils, 2008). Users either download episodes individually for listening or subscribe to the feed of a podcast, so that new episodes are automatically downloaded as they are published. Not every podcast is an equally valuable source of information and entertainment. Finding worthwhile podcasts among the large volumes of podcasts available online, which vary widely in quality and repute, can be a daunting task for podcast listeners and subscribers. We present an analysis framework, called PodCred, for assessing the credibility and quality on podcasts on the Internet. The framework is designed to support prediction of whether a listener will select one podcast over another, given that both podcasts contain comparable informational content. We demonstrate the utility of the framework with a validation exercise that demonstrates its ability to support prediction of listener appeal, i.e., the potential of a podcast to enjoy favor and preference among users.[1]

Podcasts are compared to radio programs by some definitions (Heffernan, 2005; Matthews, 2006). However, podcasting on the Internet and radio broadcasting are characterized by three main differences. First, a podcast targets a specific group of listeners who share a focused interest. The tight thematic focus of podcasts has inspired the term *narrowcasting* (Louderback, 2008). Podcasters creating podcasts anticipate longer shelf lives since it is possible to make podcasts available indefinitely for download or reuse (Louderback, 2008). Third, no specialized equipment is required to produce and publish podcasts (Geoghegan & Klass, 2005). The podosphere, the totality of all podcasts on the Internet, contains a high proportion of unscripted, unedited, user-generated content alongside professionally produced content. These characteristics of the podosphere contribute to the

Received April 28, 2009; revised September 13, 2009; accepted September 29, 2009

© 2009 ASIS&T • Published online 12 November 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21259

[1] This paper is a synthesis and extension of work presented by the authors in Tsagkias et al. (2008, 2009).

pressing need for techniques that support users in finding podcasts worth their listening time.

The task of bringing users together with podcasts they want to listen to is made challenging by the sheer number of podcasts available.[2] Download statistics reveal a steady upward trend in podcast use (Madden & Jones, 2008; van Gils, 2008). The podosphere is growing and its growth is foreseen to continue into the future (Arbitron/Edison, 2008; Matthews, 2006). Listeners require methods of discovering podcast episodes and podcasts that they would like. They need to be able to locate podcasts that treat subject material that they are interested in, an issue that has attracted recent research interest (Celma & Raimond, 2008; Ogata, Goto, & Eto, 2007). Helping listeners to find podcasts by topic is only one part of the challenge, however. Not all podcasts treating the same topic will be equally worthwhile. In this paper we address the challenge of automatically identifying which podcasts have the highest potential for listener appeal. A podcast access system can then use this information to support the podcast search and discovery process by integrating it into a ranking score or by using it to inform browsing or recommendation.

In this work, we present an approach for characterizing and exploiting the inherent properties of podcasts that signal credibility and quality to listeners. We formulate our analysis of these properties into a framework called PodCred, which consists of four categories of indicators that capture different facets contributing to listeners' acceptance and approbation. The categories contain indicators involving the *Podcast Content*, *the Podcaster*, *the Podcast Context*, and *the Technical Execution* of the podcast. The framework is aimed at providing support for the design of a system that automatically predicts listener preference for podcasts. The PodCred framework was formulated to be maximally comprehensive and independent of considerations of technical constraints on feature extraction. In this way we ensure that future evolution in automatic analysis techniques can be incorporated into systems that are based on the framework.

To validate the usefulness of the PodCred framework, we select PodCred indicators as the basis of an implementation of a basic podcast classification system. We are interested in determining whether or not indicators that can be encoded as easily extractable surface features are useful for identifying podcasts that are preferred by listeners. This basic classification system provides a foundation from which to, in the future, implement a more sophisticated system that attempts to exploit a larger range of features derived from PodCred indicators.

The PodCred framework is designed to cover a particular domain. At the most general level, that domain can be described as the podosphere, which comprises all podcasts available on the Web. The podosphere, however, can be further divided into music-based podcasts and spoken word podcasts. Our work concentrates on podcasts containing spoken content. The podosphere is not characterized by a formal genre structure, however. Rather, podcasts tend to fall into genre categories, as has been noted, for example, by Heffernan (2005). Two central genres of spoken word podcasts are particularly salient: talk show podcasts, which can also be redistributions of shows that have run on the radio, and how-to podcasts, which give commentary or advice on particular subjects. It is important to clearly differentiate podcasts from other forms of Internet multimedia, such as single audio or video files published on the Web. In particular, the following Internet multimedia sources are excluded from our domain of investigation: Viddler,[3] livestreams, Internet radio, such as Live365,[4] audio books, spoken Wikipedia articles,[5] and sites that use speech synthesis to create feeds of audio material from content originally created as text, such as Speakapedia[6] and Dixero.[7]

We foresee that as podcasting continues to mature as a form of multimedia creation and delivery, it will expand with respect to end device (for example, become more oriented to mobile phones) and/or shift medium (include increasing amounts of video content). Bloggers delight in announcing the demise of podcasting,[8,9] and often the rise of video is cited as the cause. Independently of the perceived trends in the audio-only podosphere, the phenomenon of a syndicated multimedia series that can be generated without professional equipment and which is targeted toward a specific audience is sure to endure. The framework we propose provides a fundament on which analysis of this phenomenon can be built.

We have designed the PodCred framework with the following search scenario in mind: a user makes use of a podcast search engine to search for a podcast on a particular topic with the goal of subscribing to that podcast. The search engine returns a list of podcasts in response to a user query or a request for a recommendation. The user reviews these podcasts by reading the feed-level metadata (i.e., podcast title and description) scanning the list of episodes and listening to, or briefly auditioning, a couple of the episodes. We are interested in understanding on the basis of a relatively quick review of a podcast, what motivates a user to choose to subscribe to one podcast over another.

In the next section we overview related literature. Then we discuss our motivation for analyzing podcast preference in terms of user perceptions of credibility and quality and for treating podcasts as a separate case from other types of media. Next, we present the PodCred framework and follow

[2]Apple iTunes, one of the most extensive Podcast directories, advertises an inventory of 100,000 podcasts.

[3]http://www.viddler.com/ Retrieved March 20, 2009.

[4]http://www.live365.com/ Retrieved March 20, 2009.

[5]http://en.wikipedia.org/wiki/Wikipedia:Spoken_articles Retrieved March 20, 2009.

[6]http://shinydevelopment.com/speakapedia Retrieved March 20, 2009.

[7]http://www.dixero.com/ Retrieved March 20, 2009.

[8]http://althouse.blogspot.com/2007/08/podcasting-is-dead.html Retrieved March 20, 2009.

[9]http://www.informationweek.com/blog/main/archives/2008/01/is_podcasting_d.html Retrieved March 20, 2009.

DOI: 10.1002/asi

with a validation of the PodCred framework based on a basic classification system that uses indicators from the PodCred framework that can be encoded as features that are easy to extract from surface characteristics of podcasts. Finally, we report on investigations of podcasts in the real world using our online implementation of the basic classification system. The concluding section offers a summary of our contributions and an outlook on future work.

Related Work

This paper is related to two types of literature: first, the theoretical literature on credibility and quality of information, and second, the applied literature on systems for user preference prediction. The theoretical literature on user perceptions and judgments of the quality and credibility of media will be treated in more depth in the next section, where the PodCred framework is presented. This body of literature constitutes a central underpinning of the PodCred analysis framework. Of particular importance is the literature on nontraditional media such as online content, overviewed by Metzger, Flanagin, Eyal, Lemus, and McCann (2003). Also relevant is work that has been carried out on the blogosphere, the totality of all blogs on the Web. Credibility and attractiveness of blog content involves user perceptions of the reliability of primary source information embedded in a social network (Mishne, 2007; van House, 2002) and we consider many aspects of blog preference to have relevance in our research.

The applied literature that is related to our work has been carried out in the area of user preference prediction and incorporates information about topic-independent appeal into retrieval and recommendation algorithms. In particular, researchers in the area of text-based user-generated content tackle issues of wide quality fluctuations that also pose a challenge in the podosphere. In the domain of user-contributed reviews, structural, lexical, syntactic, semantic, and metadata features have been used for automatic assessment of review helpfulness (Kim, Pantel, Chklovski, & Pennacchiotti, 2006). In the domain of online discussions, the quality of posts has been automatically assessed using a combination of features from categories with the following designations: surface, lexical, syntactic, forum specific, and similarity (Weimer, Gurevych, & Mühlhüser, 2007). Community-based answers to questions have also been automatically assessed for quality, expressed as *user satisfaction* (Agichtein, Castillo, Donato, Gionis, & Mishne, 2008; Liu, Bian, & Agichtein, 2008). In the blogosphere, research has investigated the exploitation of topic independent information for improving the quality of information retrieval. Features encoding post-level and blog-level credibility indicators have been used as (query-independent) priors to help improve blog post retrieval effectiveness (Weerkamp & de Rijke, 2008).

In the multimedia analysis community, much work has been dedicated to assessing quality of service. Of particular relevance here is the concept of *Quality of Perception* (see, e.g., Ghinea & Thomas, 2005), which emphasizes the user perspective on the technical issues of quality of service. This work recognizes the impact of topic-independent video characteristics on user satisfaction during the multimedia consumption experience. In the domain of multimedia, surface features such as length and temporal patterns have been shown to contain useful information for retrieval (Westerveld, de Vries, & Ramírez, 2006). The basic system that is implemented here in order to validate the PodCred analysis framework also seeks to exploit the contributions of surface features of podcasts to solve the challenge of predicting podcast preference. The PodCred framework is designed to be applied ultimately to the discovery of recently debuted podcasts that have the potential to become popular. For this reason, we do not include any indicators that capture the reception of the podcasts among listeners, e.g., we exclude indicators reflecting listener scores, recommendations, reviews, or download statistics. In this way, our work differs from other research on predicting the popularity of online multimedia (e.g., Szabó & Huberman, 2008).

Finally, our work has an affinity with research and development in the area of content recommendation. Recommender systems incorporate information about users for whom the recommendations are formulated, but also about the items that are recommended. Information about the perceived quality can be incorporated to refine the recommendation process (see, e.g., Adomavicius & Tuzhilin, 2005).

The PodCred Framework

Motivation for the Framework

The PodCred framework consists of a list of indicators that encode factors influencing listener perceptions of the credibility and quality of podcasts. We adopt an information science perspective and consider credibility to be a perceived characteristic of media and media sources that contributes to relevance judgments (Rieh & Danielson, 2007). Perceptions of quality and innate attractiveness are closely associated with credibility, with some work identifying quality as the superordinate concept (Hilligoss & Rieh, 2007), some viewing the two as associated with separate categories (Rieh, 2002) and some regarding quality as subordinate to credibility (Metzger et al., 2003; Metzger, 2007). We incorporate quality and attractiveness by using an extended notion of credibility that is adapted to the purposes of the podosphere.

In the context of the podosphere, it is evident that credibility alone is not sufficient to capture user preferences. Expertise and trustworthiness are conventionally considered as the two primary components contributing to user perceptions of credibility (Tseng & Fogg, 1999; Metzger et al., 2003; Rubin & Liddy, 2006; Metzger, 2007). Podcast listeners, we assume, are sensitive to these factors. In other words, users prefer podcasts published by podcasters with expertise, i.e., who are knowledgeable about the subject, and who are trustworthy, i.e., they are reliable sources of information and they have no particular motivation to deceive listeners. However, users seek out podcasts not for information alone, but also in order to be entertained. Work on assessing quality of

DOI: 10.1002/asi

perception for multimedia refers to this effect as "infotainment duality" (Ghinea & Chen, 2008). The importance of this phenomenon in the podosphere is supported by recent work suggesting that the need that prompts searchers to seek podcasts does indeed comprise both an informational *and* an entertainment component (Besser, 2008). If podcasts are a pastime, users will certainly judge podcasts according to perceived information reliability, but other factors will enter into their preference formulation as well.

In order to capture additional factors considered by users, we apply an extended view of credibility in our analysis. We explicitly incorporate acceptability aspects of podcasts—with this we mean the desirability or listener-appeal of a podcast arising from sources other than those that contribute to the believability of its propositional or declarative content. The interconnectedness of acceptability and credibility is well embodied by the use of the term "credibility" in the expression street credibility, or *street cred*. In this context, "credibility" connotes acceptance and approbation. We make use of the morpheme "cred" in the framework name as a reminder that we are using a view of credibility, where, in addition to trustworthiness and expertise, attractiveness and acceptability also play a role. Our perspective on credibility is consistent with literature that observes that the dimensions along which credibility is understood or assessed differ depending on the source that is being evaluated (Metzger et al., 2003; Rieh & Danielson, 2007).

Using perceptions of user quality alone would not be sufficient to capture the factors that cause listeners to prefer one podcast over another with comparable information content. A "PodQual" framework would be a priori unsuited to model preference in a domain where user-generated content stands on equal footing with professionally generated content. "PodQual" would lack sufficient explanatory power to cover the cases in which the low-budget living room production is preferred by listeners. In the remainder of this section we discuss the literature on user-generated media that is related to the PodCred framework.

In contrast to conventional media such as newspapers and television, content published on the Internet is not subject to vetting by professional gatekeepers (Metzger et al., 2003; Metzger, 2007). The resulting freedom and variability of expression means that analysis of Internet content can prove more challenging than analysis of conventional media. Like podcasts, blogs are characterized by a temporal dimension, with new posts being added over time. Blogs are frequently user-generated and contain primary source descriptions of people's lives and surroundings; bloggers build a tightly knit social network structure (Mishne, 2007). Bloggers are individualistic and develop their own voices (van House, 2002). The podosphere is also characterized by a high proportion of user-generated content, a social network structure, and a dominance of the voice of the individual providing testimony about personal experiences or views. The literature has applied a dedicated credibility analysis framework for blogs because information seekers do not approach blogs in the same way as they approach other forms of Web content (Rubin & Liddy, 2006; Weerkamp & de Rijke, 2008). In particular, credibility building in the blogosphere is a dynamic process characterized by exchange between bloggers and readers; revelation of real-world identities and personal details is an important part of process by which bloggers establish trust (Rubin & Liddy, 2006). In the blogosphere, trust is built by revealing bias; it is not objectivity, but rather openness about individual subjectivity that makes the key contribution (Rubin & Liddy, 2006).

Research on credibility in the blogosphere is an important source of clues for understanding the perceptions of credibility and quality in the podosphere. However, it is not possible to directly adopt a blog credibility analysis framework, such as the one presented by Rubin and Liddy (2006) for use in podcast analysis. A self-evident difference between blogs and podcasts that motivates a dedicated podcast analysis framework is the fact that the core of a podcast is its audio. For this reason audio and speech characteristics must be taken into account when analyzing podcasts. A single podcast often contains rapid crossfire conversation: such exchanges are not characteristic of blogs. Other differences are more subtle. As we will see, users searching or browsing the podosphere simultaneously seek information and entertainment. Without doubt, users also expect to be entertained by blogs. However, reading a blog is a dedicated intellectual activity that does not readily admit multitasking: the reader sits at the computer and focuses on the blog content. Users often search for podcasts, however, as listening material to accompany other activities, such as housework, commuting, or exercise. Because of this behavior, an understanding of the acceptability/appeal dimension of podcasts needs to encompass aspects designed to capture the extent to which the listener can follow the content while carrying out other activities. Additionally, blogs and podcasts are different with respect to the volume of content a user can consume. The number of podcast episodes one can listen to in a single day is substantially smaller than the number of blog posts one can read or skim. The fact that podcasts require a serious commitment of listener time leads to the result that podcasts compete directly with each other for the listener's attention: subscribing to a new podcast quite possibly means dropping an old podcast (Geoghegan & Klass, 2005).

Although much of the podosphere is user-generated, podcasting clearly remains influenced by its broadcasting heritage. We were careful to consider credibility and quality indicators for radio during the formulation of the PodCred framework. In particular, we focused on indicators reflecting well-crafted audio production. Such a parallel has also been exploited in work on blog credibility, where credible blogs have been assumed to have the same indicators as credible newspapers (Weerkamp & de Rijke, 2008).

In sum, although factors impacting perceptions of credibility and quality of conventional media and user-generated media are important for the analysis of podcasts, podcasts constitute a separate medium with its own particular dimensions. For this reason we developed a dedicated framework for the analysis of credibility and quality of podcasts.

DOI: 10.1002/asi

Presentation of the Framework

The PodCred podcast analysis framework consists of a list of indicators taken into account when assessing a podcast for credibility and appeal. The framework was formulated by synthesizing the results of three studies: a review of the credibility literature, a survey of the prescriptive guidelines written for podcasters on how to create podcasts, and a data analysis of podcasts, including both a set of listener-preferred podcasts and a set of "nonpreferred" podcasts that failed to attract listener favor. In this section the PodCred framework is presented; the contributions of each of the three studies to the formulation of the framework are described and discussed.

The PodCred framework, shown in Table 1, comprises four top-level categories of indicators. The first category, Podcast Content, deals with the quality and consistency of the intellectual content of the podcast. The purpose of this category is to capture the ability of the podcast to satisfy a particular, but yet unspecified, information need of the user. Podcast Content indicators reflect whether or not a podcast is focused on a central topic or theme. Topical focus is necessary if a podcast is to provide a good fit with a specific interest or set of interests of a listener. Also included in the Podcast Content category are indicators reflecting type, composition, and source of content. These indicators are formulated so that they can capture effects specific to user-generated content, namely, that information seekers place value on personal content (Besser, 2008). Opinions, testimonials, and recommendations can be considered personal since they arise from the experience and convictions of an individual and not via the consensus of experts or by way of social convention. The second category of indicator involves the Podcaster. It is important that the creative agent of the podcast is explicitly encoded in the framework. Both expertise and trustworthiness, two main components of credibility, imply that the information source is regarded as capable of intelligence and volition, both characteristics of a human agent. Furthermore, specifically in the case of user-generated content, credibility is built by public disclosure of personal identity and personal detail (Rubin & Liddy, 2006). Of particular importance in the Podcaster category are elements relating to the speech of the podcast. Information about the style and quality of the podcaster's speech makes it possible to capture the potential appeal of the podcaster's persona and also the basic ease-of-listening of the podcast. The third category of indicator is Podcast context. This category involves indicators that capture the network of associated information sources and social players that a podcast builds around it in order to establish its reputation. User-generated content has been described as involving a process of information exchange (Rubin & Liddy, 2006). A podcast that is tightly integrated with its information sources and with its listener group has not only a clear source of information, but it also has demonstrable impact. Additionally, user-generated content builds credibility by avoiding covert bias (Rubin & Liddy, 2006). Sponsors/stores/advertisers are included in the framework because they reveal information not only about potential bias, but also about the scope of the podcast's impact. The final category of indicators is Technical Execution. These indicators are specific to podcasts and reflect how much time and effort went in to producing the podcast.

The PodCred framework belongs to a class of credibility assessment approaches that has been called *Checklist Approaches* (Metzger, 2007). Instead of building a cognitive model of the process by which credibility is assessed by humans, such approaches aim to inventory the factors that contribute to judgments of credibility. In a strict Checklist Approach, the presence of all checklist factors would indicate maximum credibility. Here the PodCred framework takes a different tactic, leaving open two questions to be resolved when PodCred is put to use in a preference prediction system. First, it is not specified whether particular indicators are positive or negative indicators of podcast attractiveness. Rate of podcaster speech, for example, could contribute to listener preference if it is fast (implies mastery of the material) or if it slow (facilitating ease of information uptake). Second, it is not specified whether all indicators are necessary for a podcast to be attractive. For example, recommendations might make a podcast more attractive, but would not be appropriate to include in all types of podcasts.

Now that we have introduced the PodCred analysis framework, we turn to a discussion of the three studies that contributed to its formulation and to the selection of indicators for inclusion.

Derivation of the Framework

Approaches to media credibility. The extensive body of literature on media credibility assessment provides the basic skeleton for the PodCred framework. Two important streams from early research on credibility as detailed by Metzger et al. (2003) are *Message Credibility* and *Source Credibility*, and these are represented by the first two categories of the framework, Podcast Content and Podcaster. Investigation of message credibility has traditionally concerned itself with the impact of characteristics such as message structure, message content and language intensity including use of opinionated language (Metzger et al., 2003). Message source credibility research deals with assessments concerning the person or organization who generates the message. These aspects of credibility are applicable not only in the area of traditional media, but also for Internet content. Source and Content are the first two facets of judgment of information quality on the Web used in the framework of Rieh and Belkin (1998). Message credibility and source credibility overlap to a certain degree and in the PodCred framework it can also be seen that certain Podcast Content indicators could be argued to also be important Podcaster credibility indicators.

Hilligoss and Rieh (2007) present a credibility framework that can be applied across resources and across tasks. Based on a diary study using 24 participants the authors collected 12 credibility assessment types, divided into three levels, construct, heuristics, and interaction. We make use of their findings on types of credibility assessment at the heuristic and

DOI: 10.1002/asi

TABLE 1. PodCred podcast analysis framework.

Podcast Content	*Spoken content*	Podcast has a strong topical focus Appearance of (multiple) on-topic guests Participation of multiple hosts Use of field reports Contains encyclopedic/factual information Contains discussion/opinions Contains commentary/testimonial Contains recommendations/suggestions Podcaster cites sources
	Content consistency	Podcast maintains its topical focus across episodes Consistency of episode structure Presence/reliability of inter-episode references Episodes are published regularly Episodes maintain a reasonable minimum length
Podcaster	*Podcaster speech*	Fluency/lack of hesitations Speech rate Articulation/diction Accent
	Podcaster style	Use of conversational style Use of complex sentence structure Podcaster shares personal details Use of broad, creative vocabulary Use of simile Presence of affect Use of invective Use of humor Episodes are succinct
	Podcaster profile	Podcaster eponymous Podcaster credentials Podcaster affiliation Podcaster widely known outside the podosphere
Podcast Context	*Podcaster/listener interaction*	Podcaster addresses listeners directly Podcast episodes receive many comments Podcaster responds to comments and requests Podcast page or metadata contains links to related material Podcast has a forum
	Real world context	Podcast is a republished radio broadcast Makes reference to current events Podcast has a store Presence of advertisements Podcast has a sponsor Podcast displays prizes or endorsements
Technical Execution	*Production*	Signature intro/opening jingle Background music (bed) Atmospheric sound/Sound effects Editing effects (e.g., fades, transitions) Studio quality recording/no unintended background noise
	Packaging	Feed-level metadata present/complete/accurate (e.g., title, description, copyright) Episode-level metadata present/complete/accurate (e.g., title, date, authors) ID3 tags used Audio available in high quality/multiple qualities Feed has a logo; logo links to homepage Episodes presented with images
	Distribution	Simple domain name Distributed via distribution platform Podcast has portal or homepage Reliable downloading

DOI: 10.1002/asi

at the interaction level. These are the levels that are relevant for the PodCred framework, which aims to capture information that will shed light on an assessment process which is superficial and of relatively short duration, i.e., the subscribe/not subscribe decision. At the heuristics level, assessment types are media-related and source-related, corresponding to the classical components of credibility. Additionally, the heuristics level contains endorsement-based assessments. In the podcast world, a podcast enjoys endorsement when listeners accept and respond well to it. Endorsement-based criteria can be found in the Podcast Context category of the PodCred framework. Finally, the heuristics level contains aesthetics-based assessments. The corresponding characteristic of podcasts is how they sound. We add a subcategory on podcaster speech and a subcategory on podcast production to capture the impression made by the audio dimension of a podcast. These elements are designed to be the counterparts of design elements in Websites, argued by Metzger et al. (2003) to contribute to Website dynamism and in this way to impact credibility.

During the development of the PodCred framework, special attention was paid to by the work of Rubin and Liddy (2006) and van House (2002) on credibility in blogs. Blogs and podcasts share commonalities because they both are social media and contain a high portion of user-generated content. They also both have a temporal dimension, meaning that they are published in a series that unfolds over time. The Rubin and Liddy (2006) framework involves several indicators that are directly translatable from the blogosphere to the podosphere. In particular, *blogger's expertise and offline identity disclosure* is integrated into the PodCred framework as a subcategory of the Podcaster indicator category called Podcaster Profile. Next, we consider indicators related to the temporal dimension of blogs; these are listed in the Rubin and Liddy (2006) framework as timeliness and organization. In the PodCred framework aspects involving the temporal dimension are incorporated as indicators relating to whether podcasts track recent events and whether they maintain a certain level of consistency and structure. Finally, a critical aspect used in the Rubin and Liddy (2006) framework is *appeals and triggers of a personal nature*. This aspect includes the literary appeal and personal connection evoked by a blog. Parallel elements are incorporated into the PodCred framework as a subcategory of the Podcaster indicator category called Podcaster Style. Work by van House (2002) stresses the importance of the connection of online and offline blogger identities and enhancing the effect of personal voice. Parallel indicators are incorporated into the PodCred framework as "Podcaster eponymous" and "Podcaster shares personal details."

Prescriptive rules for podcasting. The PodCred framework also reflects the results of a study we carried on prescriptive guidelines that are published to help podcasters create good podcasts. Experienced podcasters understand what makes podcasts popular and what kind of podcasts listeners generally like and we incorporate this information in the PodCred framework. Our study surveyed information found at Websites focusing on helping podcasters produce better shows. A good podcast is considered to be one that promotes the popularity of the podcaster and creates a community around the show with the ultimate goal of reaching more listeners. The study identified three informative sources of information and focused on these sources. First, *Podcast Academy*,[10] a podcast containing material ranging from keynotes of podcasting conferences to interviews with guests from the podcasting domain. Second, *Podcast Underground*,[11] a podcast portal that makes information available about how to improve and enhance the content and the exposure of a podcast, including an article[12] containing comments from individual podcasters who report their personal experiences, successes, and failures while experimenting with the medium. Third, *How to Podcast*,[13] a Website providing a step-by-step guide to podcast production. The guide includes a list of key elements that should be present to make a podcast worth listening to, and also a list of guidelines for measuring success in terms of number of subscribers. The study of prescriptive podcasting guidelines provided corroboration for the inclusion of the indicators drawn from the credibility literature discussed in the previous section. We now look at what our prescriptive sources have to say about each of the indicator categories.

First, the prescriptive podcast guidelines support inclusion of Podcast Content category indicators in the PodCred framework. The guidelines stress the importance of keeping the podcast focused on one topic. Evidently, podcasters' experience underlines the importance of the narrow focus on a target audience, mentioned in the introduction as one of the major differences between a podcast and a conventional radio program. Podcasts should create a meeting point for people interested in a certain topic or a specific subgenre. A podcaster should introduce listeners to the structure of the episode, making it clear to the listeners what they can expect to hear during the show. Well-structured episodes are also reported to help in guiding the podcaster in creating a natural flow and a steady pace. Podcasters who carry out background research or prepare transcripts can more easily create the desired tightness of structure and focus within their podcast episodes. A further suggestion is to maintain a parallel structure across episodes in a podcast. A repeated structure makes the podcast feel familiar to listeners and also allows them to anticipate content. All three of the sources consulted in our study underline the importance of regularity of episode releases. Again, giving listeners the power to anticipate increases podcast loyalty. Finally, interviews with popular and well-known people in the domain are highly recommended.

Second, prescriptive podcast guidelines mention many factors that support the indicators in the Podcaster category

[10] http://www.podcastacademy.com Retrieved March 20, 2009.

[11] http://www.podcastingunderground.com Retrieved March 20, 2009.

[12] http://www.podcastingunderground.com/2007tips/ Retrieved: March 20, 2009.

[13] http://www.how-to-podcast-tutorial.com Retrieved: March 20, 2009.

DOI: 10.1002/asi

of our PodCred framework. If a show is to become popular, the podcaster should be knowledgeable and passionate about the podcast topic. The prescriptive guidelines for podcasts explicitly and emphatically recommend that podcasters share personal experiences and stories. Such sharing creates a bond between listener and podcaster. Podcasters report that building two different emotions into podcast episodes makes them more appealing, e.g., love and humor, humor and sadness. In short, our sources provide direct support for the inclusion of the indicators involving personal details, affect, and podcaster credentials in the PodCred framework.

Third, strong support for Podcast Context categories emerges from the prescriptive sources. The sources advise podcasters to stay current with the developments in the podosphere in terms of which topics are treated in other podcasts of the same domain. Podcasters should also promote interaction with listeners by reacting to comments and suggestions from their audience. Podcast guidelines advise the activation of multiple interaction channels: subscription to syndication feed (e.g., iTunes), forums, voicemails, emails, blog comments, store and donation options. Podcasters' activity and response in fora discussions and comments is crucial, since it refuels the cycle of interactivity.

Fourth, our prescriptive podcast sources provided support for the indicators in the Technical Execution category of our PodCred framework. The podcast guidelines recommend enhancing audio quality by editing the final audio, e.g., adding sound effects, cross-fades between sections, removing sentence fillers (e.g., uhm, uh), and long periods of silence. A quiet recording environment and semiprofessional microphones are suggested to minimize the background noise.

Human analysis of podcasts. The final study that contributes to the formulation of the PodCred framework is a human analysis of podcasts. Two sets of podcasts are surveyed: first, prize-winning podcasts that were taken to be representative of podcasts that enjoy high levels of user preference and, second, podcasts that fail to achieve a level of popularity in iTunes are taken to be representative of podcasts that fail to attract favor and preference. The analysis of each podcast is carried out by looking at the podcast feed, the podcast portal (if there is one), and listening to at least one, but usually several, episodes from each podcast. This process is designed to parallel our search scenario where a user examines a podcast to make a subscribe/not-subscribe decision. During the analysis we were looking for support of the indicators included in the PodCred framework and we were also on the lookout for any indicators that might not yet be incorporated in the framework. The observations made during the analysis were tabulated in a set of categories that roughly corresponds to the indicators in the PodCred framework. The counts of the podcasts in the positive and the negative categories displaying each of these indicators can be found in Table 2. Lack of complete correspondence between Table 2 and the PodCred framework in Table 1 is due to the fact that the analysis was carried out as part of the development process of the framework, as opposed to being carried out after the framework had already been developed. In the rest of this section we provide more details on the human analysis, first of the preferred and then of the "nonpreferred" podcasts.

TABLE 2. Percentage of non-preferred and preferred podcasts displaying indicators.

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

Analysis of preferred podcasts. For the data analysis we chose the prize-winning podcasts as announced in Podcast Awards[14] for 2007 to be representative of popular podcasts. People's Choice Podcast Awards are an annual contest that awards a prize to the podcast accruing the most votes.

[14]http://www.podcastawards.com Retrieved August 14, 2008.

DOI: 10.1002/asi

Voting and nomination is open to the public. Podcasts nominated for the awards must have published at least eight episodes since the beginning of May of the award year. The contest offers 22 prizes, one for each of 20 genre categories (Best Video Podcast, Best Mobile Podcast, Business, Comedy, Culture/Arts, Education, Entertainment, Food and Drink, Gaming, General, GLBT, Health/Fitness, Mature, Movies/Films, Podsafe Music, Political, Religion Inspiration, Sports, Technology/Science, and Travel) and two extra awards for People's Choice and Best Produced. The categories used in the Podcast Awards correspond roughly to iTunes main categories. For our analysis, we investigated podcasts from all categories with the exception of Video Podcast since the PodCred framework does not cover video content.

During the analysis several indicators emerged of sufficient importance to merit inclusion in the PodCred framework. First, we noticed that nearly all the podcasts surveyed use a standard opening jingle. Second, a large number have associated Websites (i.e., podcast portals). Third, many include images and links.

Additionally, we observed quite a few characteristic corroborating indicators from the literature and the prescriptive guidelines. The podcasters frequently cite their sources, either by providing Website URLs, quotes from people, or book/article excerpts. Also, although most of the time podcasters used general vocabularies, terminology from the podcasts domain of topical focus was also observed. Most of the podcasts that were analyzed contained conversational style speech. Podcasts will commonly involve two speakers; one host and one guest. However, there were frequent cases where podcasts involved multiple guests or multiple hosts. Podcasters speaking in monologue used complete sentences, but sentence fragments were common in conversations between podcasters or between podcasters and guests. The regularity of episode release ranges from two episodes per day to monthly. Some podcasts failed to respect a regular release schedule, but the majority of podcasts is published on a daily or weekly basis. All but one podcast comes with complete feed metadata. For about half of the cases, podcast-level metadata is limited to a single-sentence description. At the episode level, metadata is generally rich, with only two podcasts failing to provide episode level information. Finally, the analysis revealed that interactivity between the podcaster and the listeners is an important characteristic of good podcasting. Three-quarters of the podcasters address listeners directly. The same portion of podcasters receive a large volume of comments. Community building emerged as clearly important in the analysis, with 10 podcasts providing a forum for their listeners. Forms of podcaster response to listeners were varied, with some podcasters responding to comments directly and others giving feedback from inside a podcast episode or responding on fora.

Analysis of nonpreferred podcasts. We collected a group of podcasts lacking listener appeal by using the column headed "Popular" in iTunes. For our data analysis, we selected a set of podcasts by choosing 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 take these podcasts to be representative of the sorts of podcasts that fail to inspire listener appreciation. The analysis of this set of "nonpreferred" podcasts provided additional support for our choice of indicators. Most characteristics we observed were already included in the PodCred framework. In particular, we observed that podcasts that are not popular exhibit 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 data analysis led to the discovery of one indicator not yet included in the framework, namely, that podcast episode length tends to be short for nonpreferred podcasts. One of the cases in which podcast episodes tend to be short is when a feed is 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.

The data analysis of "nonpreferred" podcast was the final step in the formulation of the PodCred framework. The rest of this paper is devoted to discussing the validation exercise that we carried out to confirm the utility of our framework for podcast preference prediction.

Validating the PodCred Framework

In order to validate the PodCred framework, we implement a basic classification system that makes use of a select set of indicators from the framework. First, we discuss the process of selecting indicators from the PodCred framework and transforming them into features to be used in the basic system. Then, we describe the experimental setup, including the dataset used for the experiments, the experimental conditions, and the evaluation metric. Finally, we present and discuss the results of the validation experiments.

Feature Engineering for Predicting Podcast Preference

Our basic system makes use of a select set of indicators from the framework, namely, indicators that are readily accessible and have promise to be discriminative. The first step in the design and implementation is to engineer the features that will be used to perform the classification. We are interested in predicting podcast preference with the simplest possible system. For this reason, we chose to carry out our validation of the PodCred framework using a basic system with features that can be extracted with a minimum of crawling or processing effort. The basic system excludes the content of the podcast audio from consideration and uses only features that are accessible via a superficial crawl of the feed. We refer to these features as "surface features." Additionally, we are interested in investigating whether or not it is possible to extract useful features from podcasts without being required to observe the feed over time. In other words, can useful features be extracted during a single crawl that takes a "snapshot" of the feed or must the crawler return to the feed

DOI: 10.1002/asi

TABLE 3. 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 discarding any information derived from its episodes (Snapshot) or of its parts (Cumulative).

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

periodically and accumulate information about feed development from which features are extracted? We choose to look at features that fall into two categories. We define *snapshot features* as features that are associated with the podcast feed and independent of the presence of podcast episodes and enclosures. This independence guarantees that the features can be collected with a single crawl. We define *cumulative features* as features calculated from information about episodes and audio file enclosures that will possibly require multiple crawls to accumulate. A summary of all features together with a short description and an indication of type is provided in Table 3. Below, we introduce them one by one.

Snapshot features. The snapshot features that we use are derived from the PodCred framework indicator category Technical Execution. In particular, we select the indicators that deal with feed-level metadata and the feed logo. The choice of feed-level metadata was motivated by our design decision to use surface features. The use of the presence of a logo and a logo link is also consistent with our design decision to use surface features, but found additional motivation during the human analysis of podcasts. Table 2 shows that preferred and nonpreferred podcasts show sharp distinctions with respect to their use of logos and links that link the logo back to a homepage or a portal. We chose to encode six different facets of feed-level metadata: the presence of description, the length of that description, the number of authors listed in the feed, whether or not the feed specifies a copyright, the number of categories listed, and the number of keywords listed. These indicators reflect the amount of care that is invested into the production of a podcast and can potentially capture effects above and beyond those related to indicators in the Technical Execution category. For example, design of a logo and a linked homepage and inclusion of keywords and categories reflect effort invested in making the podcast findable for listeners and could effectively encode indicators included in the Podcast Context category of the PodCred framework. Recall that snapshot features encode indicators that are derived from information associated with the feed itself and not with the individual episodes or audio file enclosures. In principle, snapshot features could be extracted from a feed at the moment it debuted in the podosphere, before it has published a significant number of episodes.

Cumulative features. The cumulative features that we use are derived from the PodCred framework indicator category Technical Execution, but also from Podcast Content. From the Technical Execution category we select the indicator dealing with episode-level metadata. This indicator is encoded into features representing four facets: the number of authors reported for that episode, the proportion of episodes that contain an episode description, the average length of the

DOI: 10.1002/asi

description, and the number of episodes containing a link to an episode page. Effectively, the episode-level metadata also encodes characteristics related to indicators in the Podcast Content category, since the number of authors potentially reflects the number of podcasters hosting the podcast and the description potentially reflects the length of the episode or its topical complexity.

From the Podcast Content category we select three indicators on which to base feature derivation: "Podcast maintains its topical focus across episodes," "Episodes are published regularly," and "Episodes maintain a reasonable minimal length." We encode the topical focus of a podcast by using its *coherence score* (He, Larson, & de Rijke, 2008), a measure that reflects the level of topical clustering of the podcast episodes. The coherence score is calculated by determining the proportion of pairs of episodes in a podcast feed that can be considered to be related to each other with a similarity that exceeds a certain threshold. In order to calculate this measure, we represent each episode with its title, description, and summary, if present. The coherence score is calculated automatically using lexical features derived from these metadata elements. By using the metadata we are able to ensure that this feature remains extractable with only a surface observation of the podcast, i.e., there is no need for processing or analysis of the audio file. We encode the regularity with which a podcast is published with a Fast Fourier Transform-based measure, which is described in further detail in Tsagkias, Larson, and de Rijke (2009). We also include features that are less precise in their ability to reflect regularity, but are simpler to compute. In particular, we include a feature that requires the release period to be less than 1 week, as well as features that reflect recency and raw counts of releases. Finally, we include two features that encode podcast episode length in different ways, one which looks at the duration of the audio file as reported in the feed and one which accesses length information directly by measuring the file size of the enclosed audio episode.

In the next section we turn to a discussion of the implementation of the basic system that uses the extracted features derived from PodCred framework indicators in order to classify podcasts as to whether they are "Popular" or "Nonpopular."

Experimental Setup

The aim of the basic classification system that we implement is to validate the PodCred framework, i.e., to demonstrate whether or not the framework provides a sound basis on which to build a system that predicts listener preference for podcasts. We chose to formulate the preference prediction problem as a binary classification problem. Given a podcast, our classifier will predict whether this podcast is a "preferred" podcast or a "nonpreferred" podcast. We concentrate on investigating 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 Agichtein et al. (2008) and Liu et al. (2008).

The podcast feeds used for the experiments were those feeds listed in each of the topical categories of iTunes at the time of our crawl (late August 2008). 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 Organizations, Games and Hobby, Education, Comedy, Business, and Arts. For each category we sorted the podcast feeds in iTunes using the column labeled "Popular." We then gathered information from the 10 feeds at the top of the list and the 10 feeds at the bottom list using a crawler implemented based on the SimplePie[15] library. Feeds in non-Western languages, feeds containing video enclosures, and feeds that were unreachable were discarded. Our iTunes podcast dataset contains 250 podcast feeds with a total of 9,128 episodes with 9,185 audio enclosures. In total, the audio enclosures add up to ∼2,760 hours of audio.

Our basic system consists of a classifier that is trained to separate the podcast feeds that occurred in the top 10 "Popular" positions from those which occurred in the bottom 10 positions. The exact mechanism by which iTunes calculates "Popular" is not public knowledge,[16] but we make the assumption that it is related to the number of downloads, and, as such, reflects user preference for certain podcasts. Of the 250 podcasts yielded by our podcast crawl 148 are iTunes-Popular podcasts and 102 iTunes-Nonpopular. We do not assume that the iTunes "Popular" podcasts are the ideal representation of preferred podcasts. One factor involved is that the iTunes inventory represents only a subset of the podosphere. Although this sample is extensive, presumably, it is not completely representative, but rather biased, most probably toward high-quality podcasts. Another factor is possible interaction between podcast characteristics that are made salient by the iTunes interface and user rating behavior. In particular, it is not possible to exclude the effect that a well-designed logo tempts listeners to test and consequently to rate a podcast. The popularity ratings on iTunes is an example of a winner-take-all type market. Salganik, Dodds, and Watts (2006) demonstrate that success in such a market is only partly determined by quality. Hence, by using iTunes as ground truth we are measuring the ability of our classifier to predict emergent popularity, which is a function of the market as well as of the data. However, since we limit our experiments to podcasts at the extreme popular and the extreme nonpopular end of the spectrum, it is relatively safe to assume that the level of popularity achieved in iTunes reflects a characteristic that goes beyond lucky ascendency in a winner-take-all type rating situation. All told, the size of our iTunes podcast dataset and the fact that it is associated with ground truth based on user behavior in a real-world application are advantages that outweigh its disadvantages for the purposes of our validation exercise.

[15] http://simplepie.org

[16] iTunes declined to comment on the algorithm.

DOI: 10.1002/asi

For our validation exercise, we chose to compare a Naive Bayes classifier with a Support Vector Machine (SVM) classifier and two decision tree classifiers (J48, RandomForest)—a set representative of the state-of-the-art in classification. We make use of the implementations of these classifiers provided by the Weka toolkit (Witten & Frank, 2005). We experiment with multiple classifiers in order to confirm that our results are generally valid, i.e., not dependent on any particular approach to classification.[17]

In order to investigate whether the size of the feature set can be optimized, we employ four widely used attribute selection methods from machine learning: Correlation-based Feature Selection (CfsSubSet), χ^2, Gain Ratio, and Information Gain. CfsSubSet assesses the predictive ability of each feature individually and the degree of redundancy among them, preferring sets of features that are highly correlated with the class, but have low intercorrelation (Hall & Smith, 1998). The χ^2 method selects features that are well correlated with the two classes. Information Gain prefers features that tend to describe the dataset uniquely. Gain Ratio, similarly to Information Gain, prefers features uniquely identifying the dataset but penalizes features with wide range of values. We refer the reader to Witten and Frank (2005) for more information on feature selection.

All classification results reported are averaged over 10 runs of 10-fold cross-validation. We evaluate system performance using the precision P, recall R, and F1-score, which we report for the "Popular" and "Nonpopular" class. The F1-score is the harmonic mean of P and R, as in Equation (1), where P is the proportion of positively classified objects that were correctly classified as positive and R is the proportion of positive objects in the collection that were correctly classified as positive.

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \tag{1}$$

For determining whether the difference between the experimental system and baseline performance is statistically significant, we use the Corrected Paired-T Test (Witten & Frank, 2005).

[17]For readers not familiar with classification methods, we briefly introduce the classifiers here, and refer the reader for additional details to reference works (Manning, Raghavan, & Schütze, 2008; Witten & Frank, 2005). The Naive Bayes classifier is a probabilistic model that applies the independence assumption—features representing the objects to be classified are taken to be distributed independently of one another. A Naive Bayes classifier provides a good basis for comparison since it is simple to implement, widely used, and delivers solid performance. A Support Vector Machine is a discriminative classifier that makes use of vector representations of the objects to be classified. It defines a decision boundary within the vector space that separates objects in one class from objects in the other. The decision boundary is determined by seeking maximum separation between the boundary and the classes on either side. Decision tree classifiers sort objects into classes by learning recursive splits of the collection based on object features. Decision trees are advantageous because they isolate and exploit helpful features. The J48 algorithm makes use of information gain to grow trees and the Random Forest combines decisions of multiple trees.

Results on Predicting Podcast Preference

We report on three sets of experiments investigating the potential of surface features as listed in Table 3 for predicting podcast preference, i.e., for classifying podcasts into Popular and Nonpopular. The three sets of experiments are aimed at answering three research questions: (1) Can surface features be used to predict podcast preference? (2) Must the podcast feed be monitored over time to collect information for generating features? (3) Can the size and composition of the feature set be optimized?

In our initial set of experiments we explore the individual contribution of each feature listed in Table 3. Results of single feature classification experiments are listed in Tables 4 and 5. A classifier that assigns all podcasts to the most frequent class (Popular) achieves total recall (1.00) with a precision of 0.54, leading to an F1 score of 0.74 and is used as a baseline for comparison within our experimental setting. Notice that this baseline does not represent a point of operation outside of this setting for two reasons. First, and most obvious, the random baseline classifies every podcast as "popular," which would not be helpful information in an operational system. Second, the real-world distribution of podcasts is quite likely to lean more heavily toward the "nonpopular" end of the spectrum than the distribution of our dataset. We use the random baseline because it provides a convenient and helpful point of reference in our experimental context. Single feature classification provides improvement over the random baseline in approximately half the cases. J48 is the top-performing classifier with the Random Tree classifier and the SVM general achieving only slightly lower scores. The Naive Bayes classifier reveals itself as not particularly suited for the task, presumably due to overfitting. The feature episode_authors_count yields the strongest-performing single-feature classifiers, showing statistically significant improvement over the random baseline for all four cases. Although a classification system could be built using only one feature, its success would be completely dependent on the presence of that feature in the podcast feed. Our data analysis revealed that feeds do not always contain consistent metadata, and as such a system based on more than one feature can be expected to be more robust toward missing metadata.

With such considerations of robustness in mind, we turn to our second set of classification experiments, which compares the performance of sets of features. Tables 4 and 5 include reports of the performance of our classifiers when using *all snapshot features, all cumulative features*, and *all features combined*. The set consisting of all features combined shows a statistically significant increase over the baseline for the SVM and the Random Forest classifier, with the latter achieving peak performance of 0.81 (P: 0.78, R: 0.85). The set of all cumulative features and the set of all features combined deliver roughly comparable performance. The set of cumulative features contains 13 features and is smaller than the set of all features, which contains 21. In this respect the set of all cumulative features can be regarded as a useful optimized set.

DOI: 10.1002/asi

TABLE 4. F1, precision and recall of the positive class for Naive Bayes, Support Vector Machine (SVM), and tree classifiers (J48, and RandomForest) using a single feature, snapshot and cumulative features, and all features. Boldface indicates improvement over baseline. Statistically significant improvement (↑) or loss (↓) over the baseline is also reported.

	Naive Bayes			SVM			J48			RandomForest		
Feature	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Baseline	P: 0.59/R: 1.00/F1: 0.74											
Type: *Snapshot*												
feed_has_logo	**0.66**↑	0.96↓	**0.78**↑	**0.66**↑	0.96↓	**0.78**↑	**0.66**↑	0.96↓	**0.78**↑	**0.66**↑	0.96↓	**0.78**↑
feed_logo_linkback	0.58	0.91	0.70	0.59	0.97	0.73	0.59	1.00	0.74	0.59	0.87	0.69
feed_has_description	**0.60**	0.97↓	0.74	0.59	0.97↓	0.73	0.59	0.97	0.73	**0.60**	0.97↓	0.74
feed_descr_length	**0.70**	0.39↓	0.48↓	0.59	1.00	0.74	**0.65**	0.91	**0.76**	**0.65**↑	0.64↓	0.64↓
feed_categories_count	**0.84**↑	0.25↓	0.37↓	0.59	1.00	0.74	**0.67**	0.93	**0.78**	**0.67**↑	0.83↑	0.74
feed_keywords_count	**0.86**↑	0.20↓	0.31↓	0.59	1.00	0.74	**0.72**	0.80	**0.75**	**0.70**↑	0.64↓	0.67↓
feed_has_copyright	**0.64**↑	0.84↓	0.73	**0.64**↑	0.84↓	0.73	**0.64**	0.84	0.73	**0.64**↑	0.84↓	0.73
efed_authors_count	0.59	0.98	0.74	0.59	1.00	0.74	**0.64**	0.97	**0.77**	**0.63**↑	0.96	**0.76**
All snapshot features	**0.85**↑	0.34↓	0.47↓	**0.66**↑	0.96↓	**0.78**↑	**0.71**	0.72	0.71	**0.71**↑	0.80↓	**0.75**
Type: *Cumulative*												
feed_periodicity	0.39↓	0.28↓	0.30↓	0.59	1.00	0.74	0.59	1.00	0.74	0.59	0.75↓	0.66↓
feed_period_less1week	**0.70**↑	0.71↓	0.70	**0.70**↑	0.71↓	0.70	**0.70**	0.71	0.70	**0.70**↑	0.71↓	0.70
feed_coherence	0.59	0.89↓	0.71	0.59	1.00	0.74	0.59	1.00	0.74	0.58	0.82↓	0.68↓
episode_descr_ratio	**0.60**	0.98	0.74	0.58	0.96	0.72	0.59	0.98	0.74	0.59	0.96↓	0.74
episode_avg_descr_length	**0.60**	0.28	0.35↓	0.59	1.00	0.74	0.58	0.92	0.71	**0.61**	0.61↓	0.60↓
episode_title_has_link2page	**0.78**	0.14	0.22↓	0.59	1.00	0.74	**0.68**	0.81	0.73	**0.69**↑	0.85↓	**0.76**
episode_count	**0.92**↑	0.28↓	0.42↓	0.59	1.00	0.74	**0.78**	0.80	**0.79**	**0.79**↑	0.73↓	**0.75**
episode_authors_count	**0.67**↑	0.95↓	**0.78**↑	**0.67**↑	0.95↓	**0.79**↑	**0.67**	0.95	**0.79**	**0.67**↑	0.95↓	**0.79**↑
enclosure_count	**0.91**↑	0.27↓	0.41↓	0.59	1.00	0.74	**0.79**	0.79	**0.78**	**0.78**↑	0.74↓	**0.76**
more_2_enclosures	**0.64**↑	0.94↓	**0.76**	**0.64**↑	0.94↓	**0.76**	**0.64**	0.94	**0.76**	**0.64**↑	0.94↓	**0.76**
enclosure_past_2month	**0.89**↑	0.56↓	0.68	**0.89**↑	0.56↓	0.68	**0.89**	0.56	0.68	**0.89**↑	0.56↓	0.68
enclosure_duration_avg	0.58	0.55↓	0.55↓	0.59	1.00	0.74	0.59	0.99	0.74	**0.62**	0.79↓	0.69↓
enclosure_filesize_avg	0.59	0.95↓	0.73	0.59	1.00	0.74	0.59	1.00	0.74	**0.60**	0.65↓	0.62↓
All cumulative features	**0.88**↑	0.33↓	0.46↓	**0.79**↑	0.83↓	**0.80**↑	**0.77**	0.85	**0.81**	**0.78**↑	0.85↓	**0.81**↑
Type: *Snapshot and Cumulative combined*												
All features combined	**0.87**↑	0.39↓	0.53↓	**0.78**↑	0.83↓	**0.80**↑	**0.78**	0.80	**0.78**	**0.78**↑	0.85	**0.81**↑

The set of all snapshot features prove unable to match the performance of all cumulative features and all features combined. This suggests that the information derived from the episode and the audio enclosures of podcast feeds is important and that it is not advisable to abandon features that necessitate multiple episodes or audio enclosures for calculation and for these reasons might require protracted observation of the feed to accumulate sufficient information.

In our third and final set of classification experiments (see Tables 6 and 7), we explore whether a judicious choice of features makes it possible to reduce the number of features necessary. We investigate the performance of optimized feature sets using four automatic attribute selection methods (CfsSubset, χ^2, Gain Ratio, and Information Gain). The optimized feature sets draw features from both snapshot and cumulative feature categories. We are interested in finding features that work well together and explore a selection of feature sets created by our feature selection methods. The method χ^2, Gain Ratio, and Information Gain all return a ranked list of all input features. CfsSubSet returns a reduced feature set with no ranking information. For the first three methods we define two thresholds depending on the number of features to be included in the optimized set (Top-5 and Top-10). The feature sets are presented in Table 9.

From the results reported in Tables 6 and 7, we see that using the feature set selected by CfsSubSet we can approach the performance achieved when using all cumulative features. The CfsSubset feature set contains nine features, and is slightly smaller than the cumulative feature set. Also interesting is the fact that these features are balanced: four are snapshot features and five are cumulative features. Unsurprisingly, the Naive Bayes classifier, unable to exploit helpful features in isolation, demonstrates the greatest improvement over the baseline when feature selection techniques are applied.

Looking in greater detail at Tables 6 and 7, we observe that the performance for χ^2, Information Gain, and Gain Ratio slightly increased for the Top-10 set compared to the Top-5 set. Examination of Table 9 reveals that all three methods picked up four cumulative and one snapshot feature to form Top-5 sets. For the Top-10 sets more snapshot features were included, rendering the feature sets more equally balanced. Note that these additional snapshot features are features that demonstrate predictive ability when used in isolation, i.e.,

DOI: 10.1002/asi

TABLE 5. F1, precision and recall of the negative class for Naive Bayes, Support Vector Machine (SVM), and tree classifiers (J48, and RandomForest) using a single feature, snapshot and cumulative features, and all features. Baseline for negative class reports P: 0.00, R: 0.00, F1: 0.00.

	Naive Bayes			SVM			J48			RandomForest		
Feature	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Type: *Snapshot*												
feed_has_logo	0.83	0.28	0.41	0.83	0.28	0.41	0.83	0.28	0.41	0.83	0.28	0.41
feed_logo_linkback	0.06	0.08	0.07	0.02	0.02	0.02	0.00	0.00	0.00	0.12	0.14	0.13
feed_has_description	0.35	0.05	0.08	0.10	0.02	0.03	0.10	0.02	0.03	0.39	0.06	0.10
feed_descr_length	0.46	0.75	0.57	0.00	0.00	0.00	0.69	0.28	0.38	0.50	0.50	0.49
feed_categories_count	0.46	0.92	0.61	0.00	0.00	0.00	0.77	0.33	0.46	0.64	0.41	0.49
feed_keywords_count	0.45	0.96	0.61	0.00	0.00	0.00	0.67	0.53	0.58	0.55	0.60	0.56
feed_has_copyright	0.57	0.30	0.39	0.57	0.30	0.39	0.57	0.30	0.39	0.57	0.30	0.39
feed_authors_count	0.16	0.02	0.03	0.00	0.00	0.00	0.76	0.19	0.29	0.71	0.19	0.29
All snapshot features	0.49	0.90	0.63	0.83	0.28	0.41	0.58	0.54	0.55	0.65	0.51	0.56
Type: *Cumulative*												
feed_periodicity	0.28	0.46	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.41	0.25	0.30
feed_period_less1week	0.57	0.56	0.56	0.57	0.56	0.56	0.57	0.56	0.56	0.57	0.56	0.56
feed_coherence	0.39	0.10	0.16	0.00	0.00	0.00	0.00	0.00	0.00	0.35	0.14	0.20
episode_descr_ratio	0.35	0.04	0.07	0.06	0.04	0.04	0.03	0.00	0.01	0.32	0.04	0.07
episode_avg_descr_length	0.40	0.74	0.51	0.00	0.00	0.00	0.12	0.05	0.07	0.43	0.43	0.43
episode_title_has_link2page	0.43	0.97	0.60	0.00	0.00	0.00	0.63	0.42	0.49	0.68	0.44	0.52
episode_count	0.48	0.97	0.64	0.00	0.00	0.00	0.71	0.66	0.67	0.65	0.70	0.67
episode_authors_count	0.83	0.32	0.45	0.85	0.32	0.45	0.85	0.32	0.45	0.85	0.32	0.45
enclosure_count	0.48	0.97	0.64	0.00	0.00	0.00	0.70	0.68	0.68	0.66	0.68	0.66
more_2_enclosures	0.71	0.24	0.34	0.71	0.24	0.34	0.71	0.24	0.34	0.71	0.24	0.34
enclosure_past_2month	0.59	0.89	0.71	0.59	0.89	0.71	0.59	0.89	0.71	0.59	0.89	0.71
enclosure_duration_avg	0.36	0.43	0.39	0.00	0.00	0.00	0.00	0.00	0.00	0.49	0.29	0.35
enclosure_filesize_avg	0.23	0.04	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.43	0.37	0.39
All cumulative features	0.49	0.93	0.64	0.74	0.66	0.69	0.75	0.63	0.68	0.76	0.64	0.69
Type: *Snapshot and Cumulative combined*												
All features combined	0.51	0.91	0.65	0.73	0.64	0.67	0.70	0.66	0.67	0.76	0.64	0.69

TABLE 6. F1, precision and recall of the positive class for Naive Bayes, Support Vector Machine (SVM), and tree classifiers (J48, and RandomForest) after attribute selection using *CfsSubset*, χ^2, *Gain Ratio*, and *Information Gain*. Boldface indicates improvement in performance for the respective classifier compared to Cumulative features. Statistically significant improvement (↑) or loss (↓) over the baseline are also shown.

	Naive Bayes			SVM			J48			RandomForest		
Feature	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Baseline	P: 0.59/R: 1.00/F1: 0.74											
Snapshot features	0.85↑	**0.34**↓	**0.47**↓	0.66↑	**0.96**↓	0.78↑	0.71↑	0.72↓	0.71	0.71↑	0.80↓	0.75
Cumulative features	0.88↑	0.33↓	0.46↓	0.79↑	0.83↓	0.80↑	0.77↑	0.85↓	0.81↑	0.78↑	0.85↓	0.81↑
All features	0.87↑	**0.39**↓	**0.53**↓	0.78↑	0.83↓	0.80↑	**0.78**↑	0.80↓	0.78	0.78↑	0.85↓	0.81↑
CfsSubset	**0.89**↑	**0.36**↓	**0.50**↓	0.75↑	0.83↓	0.77	**0.80**↑	0.78↓	0.78	0.78↑	0.84↓	0.80↑
χ^2 – Top 5	**0.89**↑	**0.34**↓	**0.48**↓	**0.84**↑	0.65↓	0.71	**0.79**↑	0.77↓	0.77	0.74↑	0.81↓	0.77
χ^2 – Top 10	**0.89**↑	**0.36**↓	**0.51**↓	0.75↑	**0.84**↓	0.78	**0.80**↑	0.78↓	0.78	0.77↑	0.84↓	0.80↑
Gain Ratio – Top 5	**0.92**↑	**0.36**↓	**0.50**↓	0.72↑	**0.90**↓	0.78	**0.80**↑	0.78↓	0.79	0.78↑	0.78↓	0.77
Gain Ratio – Top 10	**0.89**↑	**0.38**↓	**0.53**↓	0.75↑	0.82↓	0.77	**0.79**↑	0.78↓	0.78	0.77↑	0.83↓	0.80↑
Information Gain – Top 5	**0.91**↑	0.31↓	0.45↓	**0.89**↑	0.59↓	0.70	**0.79**↑	0.82↓	0.80↑	0.75↑	0.79↓	0.76
Information Gain – Top 10	**0.89**↑	**0.36**↓	**0.51**↓	0.75↑	0.83↓	0.77	**0.79**↑	0.76↓	0.77	0.77↑	0.84↓	0.80↑

feed_has_logo, feed_descr_length, feed_categories_count, feed_keywords_count, feed_authors_count. The composition of the best-performing feature sets in Tables 6 and 7 is consistent with our position that a feature set consisting of both snapshot and cumulative features holds promise for good performance and also for sustaining the robustness of the classification system when confronted with feeds with no episodes or with incomplete feed metadata. Finally, we observe that feature selection also holds promise to aid design decisions about how to encode indicators from the PodCred framework into features. Some indicators translate into several potential features. For example, the PodCred indicator "Episodes are published regularly" in the category Podcast Content gives rise to both more_2_enclosures and

DOI: 10.1002/asi

TABLE 7. F1, precision and recall of the negative class for Naive Bayes, Support Vector Machine (SMO), and tree classifiers (J48, and RandomForest) after attribute selection using *CfsSubset*, χ^2, *Gain Ratio*, and *Information Gain*. Boldface indicates improvement in performance for the respective classifier compared to Cumulative features. All scores are statistically significant over the baseline (P: 0.00, R: 0.00, F: 0.00).

	Naive Bayes			SVM			J48			RandomForest		
Feature	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Snapshot features	0.49	0.90	0.63	**0.83**	0.28	0.41	0.58	0.54	0.55	0.65	0.51	0.56
Cumulative features	0.49	0.93	0.64	0.74	0.66	0.69	0.75	0.63	0.68	0.76	0.64	0.69
All features	**0.51**	0.91	**0.65**	0.73	0.64	0.67	0.70	**0.66**	0.67	0.76	0.64	0.69
CfsSubset	**0.51**	**0.94**	**0.66**	**0.76**	0.56	0.60	0.70	**0.71**	**0.70**	0.75	**0.65**	0.68
χ^2 – Top 5	**0.50**	**0.94**	**0.65**	0.64	**0.79**	0.67	0.69	**0.69**	0.68	0.69	0.57	0.61
χ^2 – Top 10	**0.51**	**0.94**	**0.66**	**0.76**	0.56	0.60	0.70	**0.70**	**0.69**	0.75	0.64	0.68
Gain Ratio – Top 5	**0.51**	**0.96**	**0.67**	**0.81**	0.46	0.55	0.71	**0.71**	**0.70**	0.69	**0.66**	0.67
Gain Ratio – Top 10	**0.51**	0.93	**0.66**	0.73	0.56	0.60	0.70	**0.69**	**0.69**	0.73	0.63	0.67
Information Gain – Top 5	0.49	**0.95**	**0.65**	0.61	**0.89**	**0.72**	0.73	**0.67**	**0.69**	0.67	0.60	0.62
Information Gain – Top 10	**0.51**	**0.94**	**0.66**	**0.76**	0.57	0.60	0.68	**0.69**	0.68	0.74	0.64	0.67

enclosure_in_past_2month. The latter was identified as useful by all feature selection methods—the fact that it is more strongly preferred suggests that it is a more effective feature for encoding the indicator.

Real-World Application of the PodCred Framework

We have seen that the PodCred framework provides a sound basis upon which to build a basic classification system capable of predicting podcast preference. In this section we report on an exploratory investigation carried out in a real-world setting. The goal of this investigation (see Table 8) is to allow us to form an impression of how a preference predictor based on the PodCred framework would behave outside of the laboratory and to gain an initial idea of the robustness of the PodCred framework in handling podcasts belonging to different genre categories.

We implemented a demonstrator that generates a preference prediction for any arbitrary podcast presented to it. The demonstrator, called podTeller,[18] accepts a URL of a podcast feed and returns a score that reflects the probability that the podcast will become popular within iTunes. Underneath the hood of podTeller is one of the configurations that emerged as a top performer during our validation experiment, namely a RandomForest classifier using the optimized CfsSubset feature set (see Table 9). The classifier is trained on the entire dataset, namely, all 250 podcasts that we collected from iTunes.

For our exploratory investigation we needed a set of podcasts occurring "in the wild," i.e., outside of the iTunes settings, and another source to identify a small set of podcasts that we could assume were popular among listeners. We chose to turn again to the winners of the People's Choice PodCast Awards, which, as previously mentioned, are selected annually by popular vote. For the human analysis of podcasts discussed in the section Human analysis of podcasts the winners from 2007 were used. Our exploratory investigation uses the winners from 2008. These two sets are not mutually exclusive, meaning that we cannot claim complete independence of the design of the PodCred framework and specific characteristics of these podcasts. However, the difference between the two sets was deemed large enough for the purpose of exploration of the behavior of the preference predictor implemented in podTeller.

Results of the investigation are reported in Table 8. The table includes the names of the podcasts, the genre category[19]

TABLE 8. PodCred framework predictions for podcasts in a real-world setting. Since these podcasts won the 2008 PodCast Awards, the system is expected to classify them into the positive class (+) rather than the negative class (−). Podcasts are shown with their genre and are grouped according to whether they are predominantly *factual* or intended for *entertainment*. Prediction and confidence score are reported for PodCred classification using CfsSubset feature set and RandomForest trained on 250 iTunes podcasts. Scores were calculated in June 2009.

Podcast	Genre	Class	Confidence Score
Group: *Factual*			
Manager Tools	Business	+	1.00
This American Life	Cultural/Arts	+	0.70
Grammar Girl	Education	+	0.90
Extralife Radio	General	+	0.90
Free Talk Live	Political	+	1.00
Daily Breakfast	Religion Inspiration	+	1.00
This Week in Tech	Technology/Science	+	0.50
WDW Radio Show	Travel	+	1.00
Group: *Entertainment*			
You Look Nice Today	Comedy	+	0.70
Mugglecast	Entertainment	+	1.00
The Instance	Gaming	+	1.00
The Signal	Movies/Films	−	0.70
Catholic Rockers	PodSafe Music	+	0.60
Feast of Fools	GLBT	+	0.80
Healthy Catholic	Health/Fitness	+	0.90
Distorted View	Mature	−	0.70

[18] http://zookma.science.uva.nl/podteller

[19] Genre categories with video podcast winners are excluded.

DOI: 10.1002/asi

TABLE 9. Feature sets derived by applying *CfsSubset*, χ^2, *Gain Ratio*, and *Information Gain* attribute selection methods. *CfsSubset* returns a list of selected attributes (★). The other methods return all attributes in descending order by their score. The score is generated by the attribute selection method and is proportional to the importance of the attribute. For χ^2, *Gain Ratio*, and *Information Gain*, two sets were created: one with the Top-5 (○), and an extended one including the Top-10 (●) attributes.

Feature	Type	Cfs-Subset	χ^2	Gain Ratio	Inf. Ratio
feed_has_logo	*Snapshot*	★	●	○	●
feed_descr_length	*Snapshot*	★	●	●	●
feed_authors_count	*Snapshot*			●	●
feed_categories_count	*Snapshot*	★	○	●	○
feed_keywords_count	*Snapshot*	★	●	●	●
episode_authors_count	*Cumulative*	★	○	○	●
episode_title_has_link2page	*Cumulative*	★	●		○
feed_period_less1week	*Cumulative*		●		
episode_count	*Cumulative*	★	○	○	○
enclosure_count	*Cumulative*	★	○	○	○
more_2_enclosures	*Cumulative*			●	
enclosure_past_2month	*Cumulative*	★	○	○	○

in which they won, and the prediction of the podTeller system. A podcast is given a positive prediction if the positive class confidence score is larger than the negative class confidence score. The table reports the predicted class for each podcast the confidence score of that class. Since a podcast must receive a large number of listener votes in order to receive an award, we expect that our system should classify award-winning podcasts into the positive class. In Table 8 it can be seen that the system predictions are largely consistent with our expectations. In order to gain an impression on the possible impact of genre on prediction results, we gather podcasts into two groups based on their content and genre. One group, marked *factual* in Table 8, contains podcasts that appear to be more information oriented and the other, marked *entertainment*, contains podcasts that appear to be amusement oriented. Note that the predictive behavior of our classifier does not differ radically for the two categories. This predictive stability suggests that a classifier implemented using features derived from indicators in the PodCred framework does not suffer from an unwanted dependence on the topic or genre category of the podcast.

Although predictions on factual and entertainment podcast are apparently quite comparable, the results in Table 8 could be interpreted as suggesting that our classifier makes less reliable predictions for entertainment than for factual podcasts. Both of the podcasts that are incorrectly classified, "The Signal" and "Distorted View," are entertainment podcasts. Moreover, on average, the confidence scores for entertainment podcasts are lower than those of factual podcasts (0.7 vs. 0.8). Since the set of podcasts involved in this experiment is limited, we want to avoid drawing any hard and fast conclusions from this apparent imbalance. However, this asymmetry does indicate that the difference between entertainment and factual podcasts may be an interesting area for future investigation. Closer examination of the two misclassified entertainment podcasts reveals that both of these podcasts have feeds in which the metadata is quite spartan, for example, their feed descriptions are rather short and they are not published using a large number of category labels. Lack of detailed metadata may, in these cases, be consistent with the specific community building strategies of these podcasts. "The Signal" is related to "Firefly," a short-lived TV series with a cult status and "Distorted View" contains mature content. It is not unimaginable that these podcasts build their following by way of "word of mouth" and that this strategy is part of the defining image they cultivate. Such a strategy would be less effective for podcasts in the factual category that have informational content to offer and whose following might depend on their visibility to viewers via search engines that need detailed metadata for effective indexing. Further investigation is necessary to determine if such a strategy is characteristic of podcasts that fall into the entertainment rather than the factual category. If podcasts created for entertainment purposes are indeed frequently crafted without readily evident indicators of their characteristics or content, it is clear that it will be necessary to include more features derived from indicators from the Podcast Content category of the PodCred framework in order for the classifier to correctly predict their popularity among listeners. In sum, a classifier built using indicators from the PodCred framework and training data drawn from iTunes demonstrates prediction behavior consistent with expectation when moved beyond the iTunes setting.

Conclusion and Outlook

We have presented the PodCred framework, designed for the analysis of factors contributing to listener assessment of the credibility and quality of podcasts. The framework consists of a list of indicators divided into four categories: Podcast Content, Podcaster, Podcast Context, and Technical Execution. Together these indicators provide comprehensive coverage of the properties of podcasts that listeners consider when they decide whether or not a podcast is worth their time and make a decision to subscribe or not to subscribe to that podcast. We have shown that the PodCred framework provides a viable basis for the prediction of podcast preference by carrying out validation experiments using a basic classification system and a dataset collected from iTunes. The experimental system was implemented using surface features that are easily extracted from podcasts. The results of the experiments demonstrate that such features can be successfully exploited to predict podcast preference, making it possible to avoid deeper processing, e.g., computationally expensive analysis of the podcast audio file. Although podcast preference can be predicted using "snapshot" information derived from a single crawl of the feed, "cumulative" information requiring repeated visits of the crawler also makes an important contribution. The best feature sets consists of a combination of feed-level and episode and enclosure-level features. An exploratory investigation of data beyond the iTunes dataset suggested that our basic classification system is capable of achieving robust performance outside of

DOI: 10.1002/asi

the laboratory and that this performance does not show signs of unduly large dependencies of classification accuracy on podcast content or genre. In total, the results of our experimentation and investigation speak strongly for the general applicability of the PodCred framework.

Future work will pursue the issue opened by our exploratory investigation of real-world application of the PodCred framework, namely the external dependencies that impact preference prediction (see preceding section). In particular, we observed behavior suggesting that the basic stability of classification across genre-based podcast groups may be subject to genre-based fluctuation. Perhaps the most useful approach is to isolate the model of user assessment of credibility and quality only partially from factors such as topic and genre. In the literature on credibility and quality, there are multiple acknowledgments of topic and genre dependencies in users' credibility perceptions. Rieh and Belkin (1998) note that it is essential to recognize the relationship between how users assess content and the informational problem they are facing. For example, medical information will be assessed in a different way from information about the personal lives of movie stars. In the former case, the information is used to make a potentially life-critical decision and in the latter case the user does not take any particular action as a result of the information. Metzger et al. (2003) observed that factual information is more rigorously checked than entertainment information. Ghinea and Thomas (2005) report that for multimedia that is educational in purpose, perceived quality does not vary widely with transmission quality. Beyond educational material, other genres do not share this stability. Future applications of the PodCred framework for the purpose of preference prediction should attempt to address the different ways in which users assess topic and genre. An adapted PodCred-based classifier could potentially avoid topic-related issues that presented a challenged for our basic classification system. For example, we observed that iTunes-Popular podcasts include examples of podcasts no longer currently publishing, but whose topic is timeless so that they do not go out of date. We observed that our basic classification system misclassified a podcast of the how-to genre on the subject of knitting, which was popular, but had no recent episodes. This example supports the perspective that recency of publication may be an important indicator of popularity for some genres, but for other genres that it is inappropriate and suggests that an appropriate extension of the basic classification system might serve to cover it.

Dependency on topic and genre can be expected to have a large user-dependent component. Users having a high level of information literacy assess credibility with different strategies. Variation introduced by users is related to user knowledge and therefore has a topic-dependent component. Users assess content in a different manner if it treats a topic that they are knowledgeable about (Metzger et al., 2003). Further, as previously mentioned, new podcasts are measured in appeal with respect to the podcasts that a user already subscribes to. A user replacing an old podcast with a new, more attractive podcast on the same subject will deploy a different assessment strategy than a user trying to increase topical coverage by introducing new podcasts on novel subjects to the subscription list. For example, we observed examples that suggest that there are groups of users whose needs are not adequately met by a system that treats all users as identical. Many false positives returned by our basic classification system were podcasts that seemed quite appealing and displayed a full range of preference indicators from the PodCred framework. These cases were often podcasts of relatively narrow appeal and of interest to a certain locality, e.g., targeted to residents of a particular city. Apparently, our basic classifier system is already capable of a basic form of discovery of podcasts with a potential for appeal. Future systems should include extensions that allow explicitly modeling specific users and specific user groups.

Finally, our initial experimentation with podcast ranking, reported in Tsagkias et al. (2009). shows that surface features derived from indicators in the PodCred framework have the potential to help predict not only which podcasts are popular, but also the degree of their popularity, as reflected by number of popularity bars assigned in iTunes. We would also like to go beyond surface features that are easily extracted using automatic methods and investigate whether additional indicators from the PodCred framework can be exploited for podcast preference prediction. For example, in order to derive features from the indicator "use of broad creative vocabulary," it would be necessary to extract statistics from a transcript of the podcast audio. This type of feature extraction is computationally more expensive, but also requires additional development effort. The use of additional indicators holds the potential to yield gains in classification performance. In sum, the PodCred framework provides a basic model of listener assessment of podcasts that can be used as a basis for the implementation of a simple classifier making use of surface features and that also offers a foundation on which to base more sophisticated podcast preference predictors in the future.

Acknowledgments

We thank Wouter Weerkamp whose expertise on blog retrieval was indispensable to our work and who contributed as a coauthor in Tsagkias et al. (2008). We benefited from the feedback from our anonymous reviewers. This research was supported by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments (http://www.stevin-tst.org) under project number STE-09-12, and by the Netherlands Organization for Scientific Research (NWO) under project numbers 017.001.190, 640.001.501, 640.002.501, 612.066.512, 612.061.814, 612.061.815, 640.004.802.

References

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Trans on Knowledge and Data Engineering, 17(6), 734–749.

DOI: 10.1002/asi

Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008). 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.

Arbitron/Edison (2008). The podcast consumer revealed 2008: The Arbitron/Edison internet and multimedia study.

Besser, J. (2008). Incorporating user search goal analysis in podcast retrieval optimization. Master's thesis, Saarbrucken, Germany: Saarland University.

Celma, I., & Raimond, Y. (2008). Zempod: A semantic web approach to podcasting. Journal of Web Semantics, 6(2), 162–169.

Geoghegan, M., & Klass, D. (2005). Podcast solutions: The complete guide to podcasting. New York: Friends of ED.

Ghinea, G., & Chen, S.Y. (2008). Measuring quality of perception in distributed multimedia: Verbalizers vs. imagers. Computers in Human Behavior, 24(4), 1317–1329.

Ghinea, G., & Thomas, J. (2005). Quality of perception: User quality of service in multimedia presentations. IEEE Transactions on Multimedia, 7(4), 786–789.

Hall, M., & Smith, L. (1998). Practical feature subset selection for machine learning. In Computer Science '98. Proceedings of the 21st Australasian Computer Science Conference ACSC'98 (pp. 181–191).

He, J., Larson, M., & de Rijke, M. (2008). Using coherence-based measures to predict query difficulty. In 30th European Conference on Information Retrieval (ECIR 2008) (pp. 689–694). Berlin: Springer.

Heffernan, V. (2005, July 22). The podcast as a new podium. The New York Times. Retrieved October 29, 2009, from http://www.nytimes.com/2005/07/22/arts/22heff.html

Hilligoss, B., & Rieh, S.Y. (2007). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. Information Processing and Management, 44(4), 1467–1484.

Kim, S.-M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 423–430).

Liu, Y., Bian, J., & Agichtein, E. (2008). Predicting information seeker satisfaction in community question answering. In SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 483–490).

Louderback, J. (2008). Master radio techniques and avoid radio traps, pnme 2007: Master radio techniques. Retrieved October 29, 2009, from http://podcastacademy.com/2008/06/18/pnme-2007-master-ratio-techniques/

Madden, M., & Jones, S. (2008). Podcast downloading 2008. Technical report, Pew Internet and American Life Project.

Manning, C.D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge, UK: Cambridge University Press.

Matthews, K. (2006). Research into podcasting technology including current and possible future uses. Electronics and Computer Science, University of Southampton, UK.

Metzger, M.J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. Journal of the American Society for Information Science and Technology, 58(13), 2078–2091.

Metzger, M.J., Flanagin, A.J., Eyal, K., Lemus, D.R., & McCann, R. (2003). Credibility in the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment (pp. 293–335). Mahwah, NJ: Lawrence Erlbaum.

Mishne, G. (2007). Applied text analytics for blogs. PhD thesis, University of Amsterdam, Netherlands.

Ogata, J., Goto, M., & Eto, K. (2007). Automatic transcription for a web 2.0 service to search podcasts. Antwerp, Belgium: Interspeech.

Patterson, L.J. (2006). The technology underlying podcasts. Computer, 39(10), 103–105.

Rieh, S.Y. (2002). Judgment of information quality and cognitive authority in the web. JASIST, 53(2), 145–161.

Rieh, S.Y., & Belkin, N.J. (1998). Understanding judgment of information quality and cognitive authority in the WWW. Journal of the American Society of Information Science and Technology, 35, 279–289.

Rieh, S.Y., & Danielson, D.R. (2007). Credibility: A multidisciplinary framework. Annual Review of Information Science and Technology, 41(1), 307–364.

Rubin, V.L., & Liddy, E.D. (2006, March). Assessing credibility of weblogs. Paper presented at the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (CAAW), Stanford, CA.

Salganik, M.J., Dodds, P.S., & Watts, D.J. (2006). Experimental study of inequality and unpredictability in an artificial cultural market. Science, 311(5762), 854–856.

Szabó, G., & Huberman, B.A. (2008). Predicting the popularity of online content. Retrieved October 29, 2009, from http://www.hpl.hp.com/research/scl/papers/predictions/

Tsagkias, M., Larson, M., Weerkamp, W., & de Rijke, M. (2008). PodCred: A framework for analyzing podcast preference. In Second Workshop on Information Credibility on the Web (WICOW 2008), Napa Valley, CA: ACM.

Tsagkias, M., Larson, M., & de Rijke, M. (2009). Exploiting surface features for the prediction of podcast preference. In 31st European Conference on Information Retrieval Research (ECIR 2009) (pp. 473–484). Berlin, Germany: Springer.

Tseng, S., & Fogg, B.J. (1999). Credibility and computing technology. Communications of the ACM, 42(5), 39–44.

van Gils, F. (2008). PodVinder: Spoken document retrieval for Dutch pod- and vodcasts. Master's thesis, University of Twente, Netherlands.

van House, N. (2002). Weblogs: Credibility and collaboration in an online world. Unpublished manuscript.

Weerkamp, W., & de Rijke, M. (2008). Credibility improves topical blog post retrieval. In Proceedings of The Annual Meeting of the Association for Computational Linguistics: Human Language Technology Conference (ACL-08:HLT) (pp. 923–931). East Stroudsburg, PA: ACL.

Weimer, M., Gurevych, I., & Mühlhüser, M. (2007). Automatically assessing the post quality in online discussions on software. In ACL 2007 Demo and Poster Sessions (pp. 125–128). East Stroudsburg, PA: ACL.

Westerveld, T., de Vries, A., & Ramírez, G. (2006). Surface features in video retrieval (pp. 180–190). In Adaptive multimedia retrieval: User, context, and feedback. Berlin/Heidelberg, Germany: Springer.

Witten, I.H., & Frank, E. (2005). Data mining: practical machine learning tools and techniques. San Francisco: Morgan Kaufmann Publishers.

DOI: 10.1002/asi