

Current Challenges and Future Directions in Podcast Information Access

Rosie Jones^{*1}, Hamed Zamani^{*2}, Markus Schedl³, Ching-Wei Chen¹, Sravana Reddy¹, Ann Clifton¹, Jussi Karlgren¹, Helia Hashemi², Aasish Pappu¹, Zahra Nazari¹, Longqi Yang⁴, Oguz Semerci¹, Hugues Bouchard¹, Ben Carterette¹

¹ Spotify ² University of Massachusetts Amherst ³ Johannes Kepler University Linz ⁴ Microsoft
¹ {rjones, cw, sravana, aclifton, jkarlgren, aasishp, zahran, oguz, hb, benjaminc}@spotify.com
² {zamani, hhashemi}@cs.umass.edu ³ markus.schedl@jku.at ⁴ longqi.yang@microsoft.com

ABSTRACT

Podcasts are spoken documents across a wide-range of genres and styles, with growing listenership across the world, and a rapidly lowering barrier to entry for both listeners and creators. The great strides in search and recommendation in research and industry have yet to see impact in the podcast space, where recommendations are still largely driven by word of mouth. In this perspective paper, we highlight the many differences between podcasts and other media, and discuss our perspective on challenges and future research directions in the domain of podcast information access.

CCS CONCEPTS

• **Information systems** → **Specialized information retrieval; Multimedia and multimodal retrieval; Speech / audio search; Summarization; Information retrieval; Recommender systems;**

KEYWORDS

podcasts, spoken documents, search, summarization, recommendation

ACM Reference Format:

Rosie Jones^{*1}, Hamed Zamani^{*2}, Markus Schedl³, Ching-Wei Chen¹, Sravana Reddy¹, Ann Clifton¹, Jussi Karlgren¹, Helia Hashemi², Aasish Pappu¹, Zahra Nazari¹, Longqi Yang⁴, Oguz Semerci¹, Hugues Bouchard¹, Ben Carterette¹. 2021. Current Challenges and Future Directions in Podcast Information Access. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, July 11–15, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3404835.3462805>

1 INTRODUCTION

With the click of a button, virtually any person with a smartphone and a podcast app such as Anchor[4] or Podbean [64] can record, edit, and publish a podcast to all of the leading audio streaming platforms. As podcasting has greatly reduced the cost of producing

and distributing audio content, there has been a massive increase in the number of podcasts being created: as of January 2021, the podcast search engine Listen Notes[58] lists over 1.9M podcast shows, and over 90M episodes hosted on public RSS servers. This has more than doubled from the end of 2019.

The listening audience for podcasts has also kept pace, growing to a critical mass in recent years. Edison Research reports[73] that podcast listening grew from 11% of the US population in 2006, to 55% in 2020. The same report found that in the US, weekly listeners spent an average time of 6 hours and 39 minutes listening to podcasts. Growth in the segment is expected to continue at a rapid pace. According to PwC[66], global podcast listenership was around 600M in 2019, and is projected to grow to 1.5B by 2024. PwC also project that podcast advertising will approach \$3.5B in 2024, an increase from roughly \$1B in 2019.

With a massive amount of podcast content, and an eager audience, there are many open questions around how best to provide access to this information. Past work in spoken document retrieval [28] is based on news corpora, while podcasts come in many disparate genres. Recommendations from friends and family [72] remains in the top-three ways people find podcasts, while non-podcast-listeners in the same study say that they don't know how to find a podcast. We believe that existing technology is insufficient for providing efficient and effective access to podcasts, which necessitates further research on the topic.

In this paper, we lay out the challenges of podcast information access, and highlight areas which are important for further research. We introduce the basic characteristics of podcasts and highlight their major similarities and differences with other media (Section 2), calling out how this leads to the challenges in information access, as well as an opportunity to use a multimodal approach to information access. We show challenges for representing podcasts for downstream information access tasks in Section 3, and highlight opportunities for further research in podcast representation. We expand upon podcast consumption patterns, listening behaviors, and potential implicit feedback signals that can be used for estimating user satisfaction for training and evaluating podcast information access systems (Section 4). We further provide an in-depth perspective on research potential related to information access technologies in Section 5. They include podcast search, recommendation, and social discovery, e.g., through social media, in addition to podcast summarization, which is a necessary component in information access systems to provide a preview of podcasts to the users. We also briefly touch on a wide range of aspects related to user experience

* Equal contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

© 2021 Copyright held by the owner/author(s). Publication rights licensed to Association for Computing Machinery.

ACM ISBN 978-1-4503-8037-9/21/07...\$15.00

<https://doi.org/10.1145/3404835.3462805>

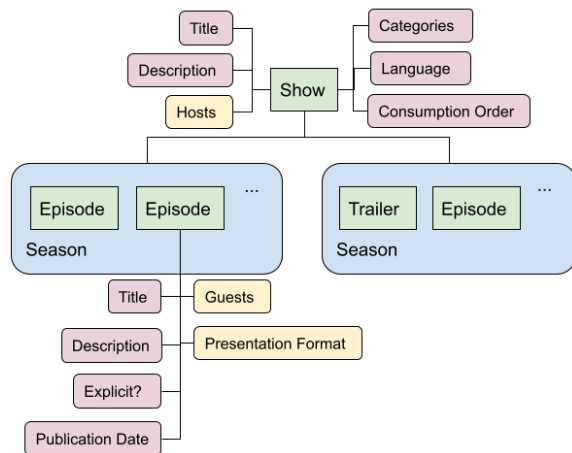


Figure 1: Structure of the metadata associated with a podcast. Metadata in pink is provided by the podcast creator in the RSS feed, while the yellow boxes are examples of properties that are not explicitly specified.

in podcast information access, which is another under-explored facet for future research.

Together, this overview will demonstrate that podcasts are significantly different from other spoken document corpora, and that we should not treat them as “noisy text” using pipelined approaches. Rather, podcasts should be handled using holistic approaches that take advantage of their multimodal and hierarchical signals. This points to a future of podcast research integrating audio and text approaches, hierarchical and end-to-end models, and representing both listeners and creators.

2 PODCAST PROPERTIES

In this section we describe the unique properties of podcasts.

Structure and metadata. Podcasts are distributed as audio streams or files, traditionally through RSS feeds. The RSS standard for podcasts contains multiple metadata fields[5]. Figure 1 illustrates the hierarchical structure of a typical podcast, and some of the metadata that is associated with shows and episodes. A podcast *show* has a title, description, language, consumption order (episodic or sequential), and a list of categories (e.g., Society & Culture, Sports, and Comedy) selected by the creator from a predefined taxonomy; the show is represented by the RSS feed. A show typically comprises multiple *episodes*, which are the distinct audio files streamed or downloaded by a listener. Each episode has its own title, description, artwork, and other information. Episodes may be organized into *seasons*, though this is generally an informal designation with no associated metadata.

As noted in previous studies [78], the metadata found in the RSS feeds is often noisy and inadequate. Episode descriptions are of varying quality and scope. Category labels cannot be considered completely reliable: categories are ill-defined, and podcast creators are incentivized to list their shows under multiple categories to maximize exposure. There is a research opportunity to identify and label categories which are meaningful to podcast listeners and automatically categorize podcasts into those categories.

Format and content. Podcasting is similar in form and content to talk radio, in that it is typically a spoken-word medium. However, the relative ease and low cost of recording and publishing means there is a great deal of variability in the specifics.

Podcast episodes have a wide range of lengths, from just a few minutes to hours. On average, each podcast episode is between half an hour and an hour in length. Figure 2a plots podcast duration in a research dataset of 100,000 podcast episodes released by Spotify [18]. The span of episode lengths reflects a variety of use cases and situations for listening events.

Podcast episodes frequently contain mixed media: music, sound effects, and archival clips, as well as the recorded narration which frames the content. Since podcasts do not require visual attention, they enable a broader set of use cases than video material, similar to other audio media such as music or broadcast radio.

Speech allows for a richer communicative channel than text, including dialectal, sociolectal, and individual variation, which all is normalized when language is written. Some podcast material is scripted, like some radio broadcasts and audiobooks. Other podcasts consist of informal and unstructured dialogue, created with the audio as the primary channel. This poses a general challenge to most existing information access systems: they are built to use a written representation, and the representation of a podcast episode in writing removes much of what characterizes it. Some of the use cases we know podcasts are created for are likely to hinge on the identification of that specific variation, e.g., material produced in a local variety of a standard language.

Presentation formats of podcasts vary widely, from monologs to multi-party conversations (see Figure 2b); from lectures and narratives to interviews, sermons, debates, and chatty conversations; from newly recorded material to historical clips; from dispassionate discourse to argumentation, jokes, or rage.

Moreover, podcast episodes are frequently anchored into a shared context between the creators and audience, and may require cultural familiarity to be fully understood. Some of these characteristics are likely to inform listening choice and selection, but information access systems today *are not* equipped to classify material in non-topical categories without editorial oversight.

The range of style, format, and language variety, and the variation exhibited by new use cases, motivates research effort to model and understand usage in more detail and to identify how the variety of material can be mapped to the variety of usage through information access tools and technology. This means developing technologies for identifying and representing the variation and leveraging that variation into useful features for information access systems, both for classification and for direct presentation to users.

3 PODCAST REPRESENTATION

How we choose to represent the information contained in podcast shows and episodes plays a key role in achieving efficient and effective information access. In this section, we describe several representations suitable for podcast information access and discuss their shortcomings, as well as opportunities for research.

Metadata. Due to the diverse nature of podcast properties discussed in Section 2, which include both structured data and free

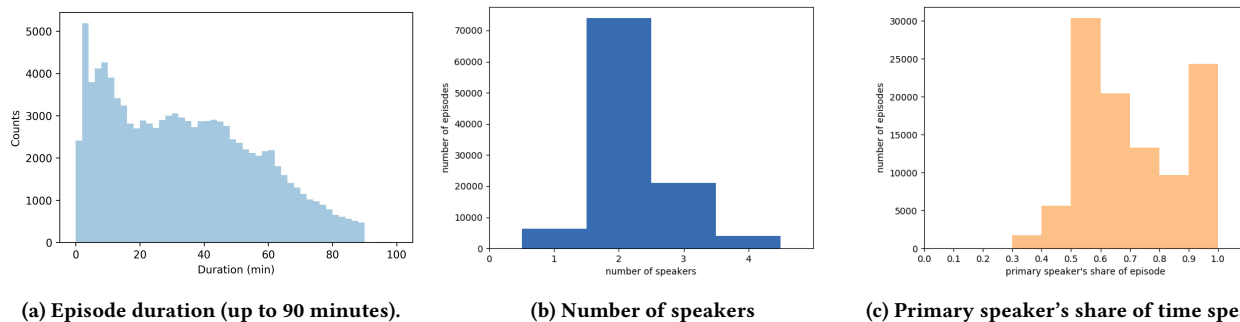


Figure 2: Distributions of episodes by duration, number of speakers, and share of primary speaker in the 100k English-language Podcast Dataset, from Clifton et al. [18]. Speaker diarization is computed automatically, and while it may be noisy, the aggregate distributions demonstrate the different conversational styles in podcasts.

text, we can use semi-structured or fielded document representations for podcasts. Free-text fields such as title and description can be treated as simple bags of words, admitting distributed vector representations.

Though optional metadata attributes and incomplete facets introduce a number of challenges for such systems, this paradigm has traditionally been quite successful in multimedia information access, such as music. However, because the essence of the podcast lies in the actual spoken audio, a deeper understanding of the podcast and its content could improve access.

There is also an opportunity to take the hierarchical nature of podcasts into account, for example, a model which can represent an episode, taking into account both the the parent show and sibling episodes, and the structure of the episode, could lead to better information access, and we recommend research in this direction.

Transcription. The main media component of podcasts is the audio stream. To enable content-based search, browsing, and recommendation, using traditional text-based approaches, a full textual representation in the form of a transcript is valuable. Human-generated transcripts are expensive and not produced by typical podcast creators. Instead, automatic speech recognition (ASR) can be used to infer a textual representation from the audio stream [97], which could then be added to the fielded document representation.

Transcribed spoken content is fundamentally different from written text due to the lack of sentential and paragraphic boundaries, as well as spoken disfluencies. Thus, spoken content is often indexed or organized differently in comparison with written text [16]. Noise due to ASR errors may be significant: the word-error-rate in the 100k Spotify Podcast Dataset is reported to be 18% [18]. Though spoken collections of news have previously been studied [23], the news domain is much more constrained and leads to better accuracy than can be expected on the wide variety of genres, levels of professionalism and languages found in podcasts. Research into ASR on the podcast domain could lead to lower error rates. In addition, using information from the podcast meta-data could lead to improved ASR performance. Another valuable research direction is end-to-end modeling, where ASR is implicitly built into the task, with audio input feeding into a deep neural model, with an information access task such as retrieval as the output.

Acoustic features. As mentioned in Section 2, podcasts may contain several kinds of non-verbal audio content. Therefore, speech transcription alone leads to information loss and thus sub-optimal information access. To address this issue, one could enrich podcast representations using acoustic features such as MFCCs [7],

PLPs [59], and more recently using ALPRs [96] that are more robust and suitable for podcasts. Such representations are effective and could leverage unlabeled data, however they are not interpretable with respect to downstream applications. We may also wish to derive interpretable features from the audio. Yang et al [96] showed they could use ALPRs to predict seriousness and energy of podcasts, as well as popularity. Acoustic features take advantage of a unique aspect of podcasts, and can be used as part of a multimodal approach to podcast information access.

Semantic podcast representation. For human-consumable podcast browsing, knowledge bootstrapping and aggregation can be useful. Semantic web techniques have been applied to podcasts to induce an RDF-like structure to the metadata and audio content [15]. A structured representation of the podcast metadata using knowledge graphs could also be computationally effective as has been found in related multimedia such as news recommendation using *NewsGraph* [45], spoken content retrieval using semantic structures [41] and large scale video classification [2].

The representation of podcasts via a heterogeneous graph could help with analysis of the most important (or well-connected) nodes, and thus their effect on downstream applications. A knowledge graph (KG) is a multi-relational, directed heterogeneous graph, composed of entities (nodes) and relations of different types (edges) [32]. KGs are used to power question answering [21], semantic search [83], and more recently recommender systems [60, 90, 91]. Because there are few explicit connections between nodes in a podcast domain, we may rely on information extraction techniques such as entity set expansion [61], relation extraction [99], and link prediction [35] to enrich the heterogeneous graph. Such a graph could be leveraged to provide insights into similarity between podcast shows and explainable connections via edges connecting their entities. One could further learn semantic relations between nodes using embedding based methods that perform random walk along the graph [14] to produce latent podcast representations.

While there are challenges in transcription, as described above, the redundancy between transcripts and descriptions could mitigate ASR errors and help us perform named entity recognition and disambiguation on the free-form text. Such entities often include people such as hosts and guests. In addition, entities such as genres, topics of discussion, and conversational styles are relevant elements of a podcast. This variety of entities could be interrelated and their interplay could result in personalized experience to the listeners.

We recommend further research into the effectiveness of existing knowledge graph work on podcast information access, as well

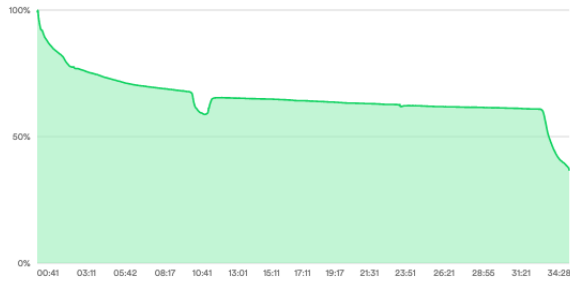


Figure 3: Illustrative curve showing the proportion of users that streamed each second of a single episode on the Spotify platform. Episodes tend to show the steepest drop-offs in listenership at the beginnings and ends, with some dips in the middle that mainly correspond to ads.

as augmenting entity extraction in a multimodal way from text descriptions and audio or transcripts of audio.

4 PODCAST CONSUMPTION AND FEEDBACK

Compared to many multimedia items, such as music, consuming podcasts requires a significant time investment. Therefore, it is particularly important to understand how users consume podcasts, as this can help us to identify implicit feedback that can be used for training and evaluation of podcast information access systems.

Podcast consumption patterns. Podcast consumption patterns are influenced by the diverse user needs they serve, as well as show characteristics such as release frequency and average episode length. Podcast consumption can be studied at an aggregated user level. A national survey conducted by Edison Research in the United States [72] revealed that the top four reasons or goals for listening to podcasts include: learning new things (74%), entertainment (71%), staying up-to-date with latest topics (60%), and getting relaxed (51%). This variety in user goals has created an ecosystem of podcast shows that are structured to address the diverse needs of the user. Informational shows that are aimed at keeping their listeners up-to-date may be released on a daily basis, while entertainment or true crime shows are often structured in seasons and released weekly, similar to TV shows. Consumption patterns in the podcast domain are different from other mediums. A study on podcast consumption patterns by Li et al. [43] demonstrated that users listen to podcasts during weekday mornings, whereas users listen to music during evenings, nights, or weekends. However, consumption patterns are personal and user level podcast consumption has been left relatively unexplored. Such analysis would lead to more accurate personalized information access systems.

Listening curves. A feature of streaming media such as audio and video is that it is possible to measure listener attention at different points in time over the duration of the stream. While tracking attention is also possible with text documents on web browsers, knowing the exact start and end points of a listener’s stream (within the client instrumentation capabilities) is a more precise signal of attention. Figure 3 is one illustrative example of a curve from listening data on Spotify that shows the proportion of listeners at each time point over the duration of a single podcast episode. In such curves, dips tend to correspond to ads or other extraneous material [68] and there are commonly sharp drops at the beginnings and the ends of episodes. These curves in general show some distinctive characteristics depending on the nature

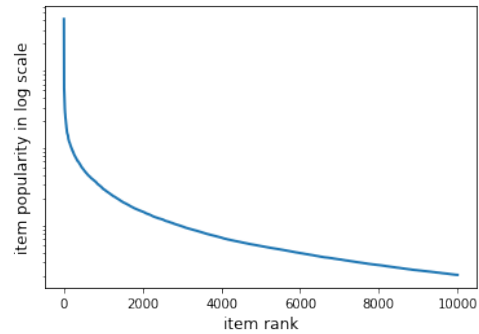


Figure 4: The long-tail characteristic of the popularity distribution of the top 10000 podcast shows on Spotify. Popularity is measured as the absolute number of streams.

of the podcast; for example, well-known podcasts tend to have a sharper drop at the beginning than lesser known podcasts, since they attract a diverse group of listeners who may be curious about the podcast but find that they are not interested after a few seconds. Listening curves are useful for detecting extraneous content[69], assessing ad monetization, improving summarization, and devising user engagement metrics on podcast-access platforms (such as, for example, deriving thresholds on the amount of listening that counts as user satisfaction).

Popularity bias. Similar to other domains such as music [38] and movies [1], a small number of podcast shows dominate the probability distribution as shown in Figure 4. Therefore, careful treatment of the items in the long-tail is necessary not only to ensure user satisfaction [81] but also to guarantee diversity and fairness [52] in podcast information access systems.

Podcast consumption as implicit feedback. On most platforms, users can subscribe to their favorite podcast shows and be notified when new episodes are released. Users may also “drop-in” to other shows and listen to specific episodes from a show due to their topic, guest or other reasons, without subscribing to the show. As in many other domains, eliciting explicit feedback from podcast listeners in scale is impractical. Therefore, inferring user satisfaction with a show or episode relies on implicit signals. In the podcast domain, *subscribing* provides a reliable form of feedback that shows the user’s interest in a show. However, it fails to create a holistic picture of a user satisfaction with the interacted item. This means that *play*, *pause*, and *stop* are also important signals, specifically on the episode level. Such signals can then be interpreted in a variety of ways to estimate user satisfaction. For instance, total listening duration, number of pauses, and listening abandonment, can be used as implicit feedback signals. Due to the high variance of podcast episode lengths, some of these signals may need to be normalized based on length of the podcast episode. Identifying and characterizing each of these signals are important research questions that need further investigation.

Moreover, having multiple implicit feedback signals provides both opportunities and challenges. At times they appear to be contradictory – users may subscribe to certain shows but not listen to them. Aggregating various signals can be challenging in inferring user satisfaction with a show or episode, and may be user-specific.

5 PODCAST INFORMATION ACCESS

Information access tools, such as search engines and recommender systems, are an essential part of finding and discovering podcasts. In this section, we highlight unique characteristics of information access in the podcast domain. We first review challenges in developing search engines and recommender systems for podcasts. We further discuss our perspective on social podcast discovery, for example through social media. We then study podcast summarization as an essential part of generating previews and trailers for podcast information access tools. We finish with covering user experiences with podcast information access systems.

5.1 Podcast Search

Podcast search shares qualities with several other search settings, while also having its own unique characteristics and challenges. In particular, podcast search is related to:

- (1) *semi-structured document retrieval*: as described in Section 3 podcasts can be represented as semi-structured documents and retrieval models, like BM25F [76], Field Relevance Models [36], and NRMF [98], can be adopted for podcast search tasks. A number of evaluation campaigns, such as INEX XML retrieval initiative [27, 39], have studied such models.
- (2) *spoken document retrieval*: podcasts can be represented by their transcripts of their spoken content - in this way podcast search is related to spoken document retrieval [28],[3],[62].
- (3) *multimedia retrieval*: as pointed out in Section 4, entertainment is one of the second most important consumption goals in the podcast domain, which is in part similar to most multimedia items, such as music and movies.
- (4) *blog search*: as argued by Besser et al. [13], the underlying goals of podcast search may be similar to those for blog search, as podcast can be viewed as audio blogs.

Below we highlight novel aspects of podcast search and potential future research directions.

Podcast search tasks. Perhaps the simplest search task in the podcast domain is *catalog search*, usually in the form of podcast show or episode title search. Misspellings, forgotten identifiers, and other errors can make the task much more difficult for a search engine to complete. For example, “tip of the tongue” search [6] is a case where a user has previously heard of or listened to an episode but cannot recall a reliable identifier.

As described in Section 4, users listen to podcasts for a variety of reasons, including entertainment, education, and information; these use cases can translate into search tasks that require more than catalog match. Podcast informational search tasks may be similar to traditional informational search, in which the user wants to find relevant information about a topic, but there are unique differences stemming from the variety of formats podcasts can take, and unique challenges due to the potential difficulty of finding relevant information in audio or noisy transcripts. Podcast segment retrieval was proposed and studied in the TREC Podcast Track [33] with informational and known-items queries. The goal was to retrieve a part of a podcast that is relevant to an information need. This is closely related to passage retrieval task with a heterogeneous collection. Traditional text retrieval approaches can be applied to

this task: for example, Clifton et al. [18] showed that term matching retrieval models, such as BM25 [75] and query likelihood [65], can achieve an NDCG@5 greater than 0.25 on a large-scale podcast search collection for a small set of queries with manual relevance annotations.

Entertainment and education tasks benefit greatly from personalized search, to leverage knowledge of the types of content users find entertaining or educational respectively. Indeed, personalized search seems to be almost a necessity for podcast search, due to varying user tastes for differing podcast formats, tolerance for low-quality audio, affinity for the hosts, not to mention contextual factors such as the window of time the user can devote to listening and what else they may be doing while they listen.

The notion of relevance in podcast search. From an information science perspective, the concept of relevance lies at the convergence of understanding users, information needs, items of information, and interaction. Relevance — the momentary quality of a text that makes it valuable enough to read — is a function of task, text characteristics, user preferences and background, situation, tool, temporal constraints, and untold other factors and has in information retrieval evaluation been formalized to be a relation between a description of a user information need and documents or information items in a collection, generalizing over other contextual or individual factors and based on topical similarity [9, 10, 54, 55, 77].

The notion of relevance for catalog search is straightforward, and for straight informational search may be a relatively straightforward translation from traditional search tasks. But because podcasts are often used simultaneously for entertainment, education, and information, the enjoyability and appeal facets of relevance are at the fore. This argues for a personalized and contextual notion of relevance. Personalization has been well-studied in web search [11, 49], and some techniques such as leveraging past consumption history are likely to be useful for podcast search as well. Contextual search has been less well-studied; work stemming from the TREC Contextual Suggestion track [19] may be the most relevant, though it focused on very specific geographical contexts.

In addition, the publication format of podcasts, as series of episodes typically consumed in sequence, and the prominence of hosts and certain popular guests, act as a filter on top of the topical search. Tsagkias et al. [87] argued that the quality and credibility of podcasts, which are sometimes considered during the relevance assessment process, can be characterized using four types of indicators pertaining to the podcast content, the podcast creator, the podcast context, or the technical execution of the podcast. These facts distinguish podcast search from most well-established search tasks including adhoc, web, personal, and enterprise search.

Podcast collections. A podcast catalog consists of podcast metadata (RSS feeds) and audio and is constantly growing as new shows and episodes are added. A podcast search collection is likely to be a snapshot of a catalog, constructed from the metadata and audio (likely via transcription). As pointed out in Section 3, transcribing podcasts is not flawless and this may influence the retrieval performance. In addition, podcasts are often long, thus their transcriptions result in very long documents. For example, the average and maximum document length in the Spotify Podcast Dataset [18]

are 5,728 and 43,504 words, respectively. It is well-known that great variation in length creates challenges for standard IR models [79].

Moreover, podcast collections are heterogeneous. Some podcasts have one speaker, while some include conversation between multiple persons. In addition, podcasts may contain non-verbal information, such as music or background audio effect. How to incorporate heterogeneous, multimodal information into a search engine is potentially a rich line of research.

Additionally, podcast catalogs are dynamic. They quickly evolve and grow — and sometimes shrink when creators delete old episodes. This calls for research on temporality and sequentiality in search. In particular, understanding the relation between world events and new items in the catalog could be beneficial in podcast search tasks.

Since podcasts are user-generated, they are of varying quality and credibility. This fact should be considered in podcast search engine development. We cover the concept of trust and credibility in Section 5.2.

Podcast search evaluation. Evaluating podcast search engines is challenging. Even after deciding on a definition of relevance, podcasts can be difficult to assess. Whether assessment is done by listening or by reading transcripts, it is potentially expensive and time-consuming. Long podcasts may touch on a diverse set of topics and therefore require careful attention to determine relevance, particularly in errorful transcripts. The TREC Podcast track [33] addressed this by making the retrieval unit two-minute segments, so that each segment was relatively easy to assess by reading the transcript, and the relevance judgements could be re-used.

Podcast relevance is often personal and contextual, which makes reusable assessments far more difficult to collect. Even when implicit feedback such as streams or subscriptions is available, it does not necessarily indicate relevance. Inferring the relationship between context, relevance, and observed implicit feedback in podcast search is challenging and requires further investigation.

With relevance assessments, traditional IR evaluation measures like precision, recall, and nDCG can be used for evaluating podcast search. Here too there is opportunity for novel research on new metrics. For example, metrics that account for the amount of time it takes to consume retrieved content could prove useful [80].

Result presentation. Search engines often provide a short preview or summary of the retrieved items for quick assessment by users, so they can faster find the relevant items. For instance, snippets are used in web search for summarizing documents. In the context of podcasts, episode descriptions can be employed as a summary, however, as mentioned in Section 2, this field is optional and of varying quality. Alternatively, search engines could display a transcription of a relevant snippet of the audio, or more generally, automatically generate a query-biased summary of the episode. This is challenging; we discuss in more detail in Section 5.4.

5.2 Podcast Recommendation

Podcast recommender systems have been categorized as speech recommenders [20], which falls short of capturing their multimodal nature discussed in Section 2. While generic recommender system

algorithms, such as variants of collaborative filtering, are applicable to this domain, the specific nature of podcasts—in particular their representation as (noisy) text through automatic speech recognition—makes content-based and hybrid approaches appear particularly well-suited. In fact, the few published approaches to podcast recommendation typically leverage textual information attached to the audio [93] or extracted through conversational interfaces [94]. A very recent approach is trajectory-based podcast recommendation, which models short-term listening behavior of users as trajectory in a podcast graph, and predicts the next shows a user is likely to access [12]. More precisely, this sequential approach represents a given collection of podcasts as a graph where nodes are shows and edges connect them through shared topics. Enriched with podcast descriptions (e.g., from Wikipedia), a graph embedding approach is applied, which represents nodes in a semantic embedding space. Users are then modeled as temporal sequences of these node embeddings, and a recurrent neural network architecture (LSTM-based) is used to predict the most likely next show the target user may listen to. While existing work has yielded interesting insights and first results, major open questions and challenges need to be addressed and investigated in depth.

Cross-domain recommendation. Since it is often music streaming platforms that extend their catalogs to include podcasts, a natural choice to address the cold-start problem (in this case, missing initial user-podcast interactions by existing users), is to adopt a cross-domain recommendation approach. In particular, using music preferences in a cross-domain fashion to address cold-start in podcast recommendation has been shown to be successful [56]. On the other hand, domains such as movies or books have not yet been investigated for cross-domain podcast recommendation. Neither have microblogs or other (textual) user-generated content shared on social media, which presumably hold rich information about topical preferences of users.

Duration-aware recommendation. Compared to music, where individual recordings have a duration of typically a few minutes, the duration of episodes in podcasts spans a wide range, from a few minutes to several hours (see Figure 2a). Furthermore, while it typically takes a listener only a couple of seconds to decide whether or not they like a song, this time is much longer for podcasts, in fact may be closer to the time needed for a movie or TV show. Poor podcast recommendations are therefore likely to more severely deteriorate user experience than poor recommendations in the music domain. The implications these characteristics have on the ways users interact with items, and consequently the requirements of a podcast recommender system, are still open to investigation.

Tailoring recommendations to situational context. Unlike music, which is frequently consumed in the background, podcasts are almost exclusively consumed in active listening modes, i.e., listeners pay close attention to the content. In addition, podcasts are often consumed while commuting [57]. As a result, temporal constraints are typically more pronounced for podcasts than for music. Because podcasts are much longer than songs (see Fig. 2a) and users tend to prefer to listen to them sequentially, both situation and timing need consideration. A user on a 30-minute-commute may prefer podcasts of approximately this duration over significantly shorter

or longer recommendations. The influence of the listeners' situational context and temporal constraints while consuming podcasts on user behavior and preferences, as well as their implications on algorithm design and system evaluation has not been investigated yet - we recommend further research in this area. Nor is it known to what extent standard context-aware recommendation algorithms can be used or need adaptation to suit the domain of podcasts.

User characteristics and podcast preferences. Extensive research has already been conducted to uncover relationships between a wide range of user characteristics and preferences for various kinds of media items; the former including user demographics [24, 71], inclination to mainstream vs. niche items [8, 89], and personality traits [22, 53]; the latter including movies [30], books [70], and music [25]. User models encoding such relationships can be integrated into recommendation algorithms and used to tailor recommendations to certain user groups, to mitigate cold start, or to diversify recommendations. Due to the recency of the podcast recommendation task, similar studies on how user characteristics shape preferences for podcasts do not exist yet. Also, we do not know yet whether inferred user models are capable of improving podcast recommendation performance similar to other domains. While a first study found that a podcast recommendation approach performs better when adopting a pre-filtering strategy with respect to the user's age [12], more in-depth investigations of user models, personalization strategies, and their impact on podcast recommendation performance are required to assess the potential of insights (and their formalization) gained in the studies mentioned.

Trust and credibility. While many podcasts are consumed as a form of entertainment, informational or educational podcasts are common, and may contain opinion or commentary on current affairs. Because of this, it is important to consider the trustworthiness of different sources, and the safety of the listener. The concept of trust and credibility [26, 29] has been studied in information retrieval, though much of that research has focused on the perceived trustworthiness of the information source [46]. More recent research has focused on online misinformation [67] and identification of fake news [40, 100]. The credibility of podcasts in particular has been studied by Tsagkias et al. [87], analyzing the factors which affect listener's perception of podcasts' credibility, including characteristics such as production quality and speaker style. As podcasts become a more popular source of information on current events, concerns over misinformation and "echo chambers" become more important. For instance, to what extent do podcast recommender algorithms reduce or amplify misinformation? Do podcast services have the ability to identify and act upon content that contains misinformation? How can podcast recommender systems reduce the risk of creating "echo chambers" which guide listeners further down paths of potentially dangerous misinformation?

Preference elicitation and evaluation. Podcasts are serial, periodic media since new episodes are released on a recurring basis. Therefore, users interact differently with podcasts than with other media types such as music or movies. As a result, implicit preference feedback differs. In particular, users can subscribe to shows and follow podcast creators, indicating they want to know when the next episode is released or their favorite creator creates a new show

or is featured in an episode, respectively. Whether a user subscribes to a podcast feed has been found to depend on a variety of factors, some of the most important being the length of the description, keyword count, whether the feed has a logo, episode duration, author count, and feed period [88]. Feedback is also available at the level of individual episodes, for example starting, pausing, skipping after a certain time. All these different kinds of feedback on (inter-related) items make interpretation of preference signals a challenge, and may even yield contradicting preference indications, for instance if someone likes a particular episode of a show or a creator appearing in an episode, but not the show in general.

These particularities of implicit feedback signals call for revisiting evaluation metrics and discussing whether we should adapt existing or devise new metrics that consider the different interaction types (e.g., subscribing, following, watching) and levels of feedback (e.g., creator, show, season, episode). Another aspect relevant to evaluating podcast recommendation stems from the fact that listening to podcasts is more time consuming than, for instance, reading a tweet or listening to a song (see above). Therefore, the negative effect of a poor recommendation on user experience and retention is higher than for many other recommendation domains. Evaluation approaches could consider this, e.g., by giving higher priority to maximize precision and minimize false positive rate.

5.3 Social Podcast Discovery

Like news readers [74], podcast listeners find content to consume outside of the context of algorithmic search and recommendation. While searching the internet is the most common way to find podcasts (77% of podcasts listeners do this occasionally), 67% of listeners find podcasts through social media posts, and 66% find out about them through recommendations from friends and family [72]. Other approaches to podcast discovery include non-podcast platform advertisements (web search/TV/radio ads), as well as cross-podcast recommendation (62%) and advertising (54%). Here we focus on social media discovery, due to its research potential.

We categorize social media discovery into two types. The first is when a listener's interest in a podcast is triggered by another user in a social media sharing information about a podcast. Since podcasts involve significant investments in time and energy, users need transparency about why a podcast is shared, to engage with the podcast. Platforms should allow users to be more specific about what they find attractive about a podcast, by enabling them to share a particular quote from a podcast, its summary, a snapshot of a conversation, etc. to be able to interest more users to engage with the podcast. In Section 5.4 we elaborate on research directions for summarization and trailer generation.

The second category is a system making a recommendation to a user based on their trust network's preferences, also known as social recommendation [34]. King et al. [37] referred to social recommendation as any recommendation with online social relations as an additional input. Social relations can be interpreted as trusted relations, friends, or followers [85]. According to this definition, social recommendation systems assume that user preferences are correlated if they establish social relations. This is in contrast to traditional recommender systems which assume users are independent and identically distributed (i.i.d. assumption) [47]. This assumption

also makes sense intuitively. In the physical world, users often seek recommendations from friends, family, and generally their trust network. Weng et al. [92] shows this is also the case in social media. The authors have shown that users with *follow* relationship are more likely to share similar interests compare to two random users.

The heterogeneous nature of social media means that different types of social relationships may have different impacts on social recommendation systems. For example user u might trust user u' about computer science, but not political topics. Using the trust relation that benefits social recommendation plays a particularly vital role when it comes to podcast media, considering their multi-faceted nature. Podcasts could have different providers, in different styles (e.g., interviews, storytelling, etc), with different hosts/guests. Even the podcast topic might change during a podcast show. All these new angles make podcasts intrinsically complex mediums. As a result, there are many problems that need to be revisited in modern sociology such as online trust [48, 84], community detection [42, 86], and heterogeneous networks [82] in podcast recommendation.

5.4 Podcast Summarization

Given that podcasts can be 30 minutes or longer per episode and the opening content does not always describe what is to come, listeners have to make an informed decision on whether the podcast is worth their time. Surveys reveal that listeners pay particular attention to the text description of a podcast in deciding whether to listen [51]. The production of audio trailers is also a way for listeners to get a preview of the podcast. The composition of informative, accurate, and catchy descriptions and trailers is a time-consuming task; many podcasts have brief, uninformative descriptions and most have no audio trailers. We see an opportunity for automatic summarization of podcasts to serve information access needs in this domain, analogous to the role of summarization of long text documents such as news stories and research literature today.

Podcast summarization is a new research area that recently got its start through the TREC 2020 Podcast Track [33]. There are known challenges presented by speech for summarization that apply to podcasts: the ambiguity of utterance boundaries, natural conversational disfluencies, lack of explicit formatting, and errors introduced by automatic speech recognition [50]. It also has unique challenges that make it distinct from traditional text or speech summarization, outlined below.

Abstractive versus extractive summarization. Because of the errors and disfluencies in transcribed speech, an abstractive model is best suited for a written text summary (although such a model may contain extractive components). Indeed, high-performing systems in the TREC 2020 podcast summarization task all produced abstractive text summaries [33]. On the other hand, a model to generate audio trailers could be primarily extractive; ideally, such a model should pay attention not only to the transcribed text content but also audio cues to indicate inclusion in a trailer. Trailer generation may also benefit from abstractive models that produce text to be read by voice actors or text-to-speech systems.

Multimodality. The generation of text summaries from podcast audio files presents a challenge, which can be addressed either through pipelined approaches that first use automatic transcription

to convert the audio to transcribed text, and then transcription to summary, or through approaches that integrate the audio signal into the summarization model. While spoken language summarization has previously been studied [17], podcasts present additional challenges in the spoken domain. Above the previously seen challenges of noisy speech transcription and sentence segmentation, podcast often contain rapid, casual speech that compound these problems. Further, podcasts are often voiced by many speakers without well-defined turn taking, and make use of nonlinguistic audio cues that are ignored by transcription systems. Future approaches towards podcast summarization should leverage the audio signal directly to avoid spuriously impoverishing the model and propagating transcription error.

Genre and use case diversity. A summarization system for podcasts must be able to handle a full array of styles and formats described in Section 2 and be robust to differences in speaking style, clarity, and structure of the content. In addition, podcast summarization calls for robustness to the style and use case of the summary. Traditional document summaries are designed to capture the key information of the documents such that a reader does not need to delve into the documents themselves if they are simply seeking to learn a few high-level takeaways. Podcasts, on the other hand, may be focused on non-informational subject matter that is more difficult to quantify informatively. Accordingly, the role of summaries for podcasts is less clear cut, and may be informational or promotional. Most podcasts are designed to be consumed as audio experiences in their full forms, with factual information being no more important than the opinions, arguments, chit-chat, music, and creative structuring of the podcast.

Many podcasts have promotional summaries, with catchy lines or hooks to entice listeners without giving away too much of the content. On the other hand, some use cases for informational summarization may call for the generation of an outline-style summary. Users may wish to consume particular segments within a podcast such as a question and answer, debate, or interview portion or other structural feature. This differs from the search task in that it is not query based but rather assumes the discovery of user- and query-independent cohesive segments within an episode than may be defined by structural or format properties rather than by topic. This type of summary should ideally automatically detect the chapter boundaries corresponding to the segment changes within a podcast.

Systems to produce automatic summaries must make deliberate choices about the role of a summary. While supervised models trained on a representative set of examples may implicitly learn to generate genre-appropriate summaries, they are not guaranteed to do so. These are considerations that should be accounted for when designing a general purpose podcast summarization system.

Contextualization. For podcasts that are serialized, with each episode building upon the previous ones, creators may choose to include a recap of the previous episodes to establish context. Summarization systems could be designed for this use case, either to aid creators in composing such recaps or exposing automatically generated summaries of previous episodes to listeners. Such systems should take into account not only the episode being recapped, but the context of the episode which the summary accompanies.

Another type of contextualization is producing personalized summaries which are tailored to a listener’s preferences, history, or a specific query. Personalized summarization has been explored in other media [31, 44] but is an unexplored avenue for podcasts.

Evaluation. Evaluation presents a significant challenge for podcast summarization. Already, there can be any number of ways to summarize a text, and generating even a single high-quality reference summary is expensive. These issues are exacerbated in podcast summarization, because of the wide variety of podcast genres and summary use cases, making it more difficult to define and quantify quality. This is compounded by the exaggerated compression ratio of podcasts to summaries in comparison to the traditional summarization task, since traditional task documents, e.g. news articles, are typically much closer in length to their reference summaries than podcasts. Initial results [18, 33] indicate that ROUGE metrics using podcast episode descriptions as the reference correlate weakly with expert human judgements, but future work should examine this more thoroughly, because the highly subjective and generative nature of producing summaries for podcasts is likely to exacerbate known issues with lexical matching metrics such as ROUGE [63]. Additionally, research into reference-free task-based evaluation will be valuable for podcast summarization, particularly in light of the varied summarization use cases for the podcast modality.

5.5 User Experience

Improving the user experience of podcast information access poses unique challenges and opportunities. In this section, we highlight four distinct aspects that future research should invest in: information access interfaces, fine-grained information access, intentional information access, and multi-sided markets.

Information access interfaces. People access podcasts mostly through visual interfaces on mobile and desktop. However, the increased popularity of voice interfaces on smart devices provides new channels for access, especially under hands-free scenarios like driving, cooking, etc. As discussed in Section 5, podcast search and recommendation additionally rely on rich metadata which is presented to the user to help them select from a range of options. When giving the listener the results of search or recommendation through a narrow voice channel, it can be inefficient to deliver and navigate through the same meta-data [94]. Future research can develop and leverage audio summarization (as also explored in Section 5.4) and text-to-speech methods to synthesize verbose information, and may also explore emerging hybrid interfaces (e.g., car displays) and multi-device settings.

Fine-grained information access. People find and discover podcasts at multiple granularity levels—a show, an episode, or a snippet within an episode. Finer-grained discovery presents new research challenges, including modeling nuanced user interactions within an episode and understanding the surrounding context in which the discovery happens: while some snippets can be consumed standalone, others may require additional background knowledge (e.g., news) or consumption order (e.g., true crime). Experimentation on retrieval of two-minute segments takes us a step in this direction [33], but future research should look at variable length segments

as has been done on news documents [23]. The user experience of snippets is an important area for future research as well.

Intentional information access. Podcast consumption has been based on RSS subscription—people subscribe to the shows they plan to listen to and then consume the released episodes regularly. This characteristic makes podcast information access intentional. A field study [95] has shown that overlooking listener intentions in podcast recommendations discourages listeners from achieving their aspirations and results in lower satisfaction. Future work should investigate mechanisms to elicit or infer intentions in a light-weight way and incorporate such information into the discovery service.

A multi-sided ecosystem. There are multiple sides in information access markets [52], as both listeners and creators have a stake in how items are ranked and recommended. Podcasts also have advertisers as stake-holders, since podcasts frequently contain advertisements, both read by the podcast host, as well as those produced by the advertiser and inserted into the audio stream. Podcast information access technology, is mostly used by listeners. However, such technology can also provide useful suggestions, in the form of metrics and consumption patterns, to creators, to help them improve their content or reach a broader audience, as well as feedback for advertisers. This introduces some interesting challenges, such as satisfaction and fairness on all sides in information access. In the podcast context, research is needed to understand the trade-offs for multiple parties. Future research can also explore designs that allow creators to customize their content for different interface (e.g., providing summaries for voice interface) and annotate their episodes for snippet-level podcast preview and listener matching.

6 SUMMARY

Our recommendations on challenges and open directions in podcast research are focused around the need to reexamine typical textual methods for the podcast domain and its novel use cases, and to leverage multiple channels of information. This includes intra- and inter-podcast structural organization and metadata, user information and listening behavior, and both linguistic and paralinguistic audio features of podcast content. In particular, for podcast **representation**, future work should focus on developing unified representations over these multiple channels, for both generic and task-specific learning. Podcast **consumption and feedback** has unique and challenging characteristics that should make it an area of focus. For podcast **search**, research should account for the wide variety of podcast consumption goals to develop appropriate notions of relevance and personalization. Podcast **recommendation** calls for cross-domain approaches, as well as duration-aware methods that leverage user context. The listening investment required of users by podcasts calls for the investigation into the process of **social podcast discovery**, via networks and their specific technologies. For podcast **summarization**, future work should be robust to the wide variety of genres and use cases, both in modeling and evaluation, and should recruit multimodal information. Finally, we advocate for research into **user experience for podcast information access**, to gain understanding into user and creator needs in different interfaces and interaction types and their impact on experience.

7 ACKNOWLEDGEMENTS

This work was supported in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

REFERENCES

- [1] Himan Abdollahpouri, Masoud Mansoury, Robin Burke, and Bamshad Mobasher. 2019. The Unfairness of Popularity Bias in Recommendation. *arXiv preprint arXiv:1907.13286* (2019).
- [2] Sami Abu-El-Hajja, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675* (2016).
- [3] Tomoyosi Akiba, Hiromitsu Nishizaki, Kiyooki Aikawa, Xinhui Hu, Yoshiaki Itoh, Tatsuya Kawahara, Seiichi Nakagawa, Hiroaki Nanjo, and Yoichi Yamashita. 2016. Overview of the NTCIR-10 SpokenDoc-2 Task. In *Proc. NTCIR-10*.
- [4] Anchor. 2021. The easiest way to make a podcast. available at <https://anchor.fm/> (accessed Feb. 9, 2021) (2021).
- [5] Apple. 2020. A Podcaster's Guide to RSS. https://help.apple.com/itc/podcasts_connect/#/itcb54353390. (2020). (Accessed on 01/30/2021).
- [6] Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. Tip of the Tongue Known-Item Retrieval: A Case Study in Movie Identification. In *Proceedings of the Sixth ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21)*.
- [7] Mathieu Barthet, Steven Hargreaves, and Mark Sandler. 2010. Speech/music discrimination in audio podcast using structural segmentation and timbre recognition. In *International Symposium on Computer Music Modeling and Retrieval*. Springer, 138–162.
- [8] Christine Bauer and Markus Schedl. 2019. Global and country-specific mainstreamness measures: Definitions, analysis, and usage for improving personalized music recommendation systems. *PLOS ONE* 14, 6 (06 2019), 1–36. DOI: <http://dx.doi.org/10.1371/journal.pone.0217389>
- [9] Nicholas J Belkin, Charles LA Clarke, Ning Gao, Jaap Kamps, and Jussi Karlgren. 2011. *Proceedings from the SIGIR workshop on "entertain me" supporting complex search tasks*. ACM New York, NY, USA.
- [10] Nicholas J Belkin, Charles LA Clarke, Ning Gao, Jaap Kamps, and Jussi Karlgren. 2012. Report on the SIGIR workshop on "entertain me" supporting complex search tasks. In *ACM SIGIR Forum*, Vol. 45. ACM New York, NY, USA, 51–59.
- [11] Paul N. Bennett, Ryan W. White, Wei Chu, Susan T. Dumais, Peter Bailey, Fedor Borisov, and Xiaoyuan Cui. 2012. Modeling the Impact of Short- and Long-term Behavior on Search Personalization. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. ACM, New York, NY, USA, 185–194. DOI: <http://dx.doi.org/10.1145/2348283.2348312>
- [12] Greg Benton, Ghazal Fazelnia, Alice Wang, and Ben Carterette. 2020. Trajectory Based Podcast Recommendation. *CoRR abs/2009.03859* (2020). arXiv:2009.03859 <https://arxiv.org/abs/2009.03859>
- [13] Jana Besser, Katja Hofmann, and Martha A. Larson. 2008. An Exploratory Study of User Goals and Strategies in Podcast Search. In *LWA 2008 - Workshop-Woche: Lernen, Wissen & Adaptivität, Würzburg, Deutschland, 6.-8. Oktober 2008, Proceedings (Technical Report)*, Joachim Baumeister and Martin Atzmüller (Eds.), Vol. 448. Department of Computer Science, University of Würzburg, Germany, 27–34.
- [14] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013), 2787–2795.
- [15] Oscar Celma and Yves Raimond. 2008. Zempod: A semantic web approach to podcasting. *Journal of Web Semantics* 6, 2 (2008), 162–169.
- [16] Ciprian Chelba, Timothy J Hazen, and Murat Saraclar. 2008. Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine* 25, 3 (2008), 39–49.
- [17] Kuan-Yu Chen, Shih-Hung Liu, Berlin Chen, and Hsin-Min Wang. 2016. Improved Spoken Document Summarization with Coverage Modeling Techniques. *CoRR abs/1601.05194* (2016). arXiv:1601.05194 <http://arxiv.org/abs/1601.05194>
- [18] Ann Clifton, Sravana Reddy, Yongze Yu, Aasish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth J. F. Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. 100,000 Podcasts: A Spoken English Document Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 5903–5917. <https://www.aclweb.org/anthology/2020.coling-main.519/>
- [19] Adrian Dean-Hall, Charles L.A. Clarke, Jaap Kamps, Paul Thomas, and Ellen Voorhees. 2012. Overview of the TREC 2012 Contextual Suggestion Track. In *Proceedings of TREC 2012*.
- [20] Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender Systems Leveraging Multimedia Content. *ACM Comput. Surv.* 53, 5, Article 106 (Sept. 2020), 38 pages. DOI: <http://dx.doi.org/10.1145/3407190>
- [21] Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 260–269.
- [22] Peter Gregory Dunn, Boris de Ruyter, and Don G. Bouwhuis. 2012. Toward a better understanding of the relation between music preference, listening behavior, and personality. *Psychology of Music* 40, 4 (2012), 411–428. DOI: <http://dx.doi.org/10.1177/0305735610388897>
- [23] Maria Eskevich, Walid Magdy, and Gareth JF Jones. 2012. New metrics for meaningful evaluation of informally structured speech retrieval. In *European Conference on Information Retrieval*. Springer, 170–181.
- [24] Bruce Ferwerda, Marko Tkalcić, and Markus Schedl. 2017. Personality Traits and Music Genre Preferences: How Music Taste Varies Over Age Groups. In *Proceedings of the 1st Workshop on Temporal Reasoning in Recommender Systems co-located with 11th International Conference on Recommender Systems (RecSys 2017), Como, Italy, August 27-31, 2017 (CEUR Workshop Proceedings)*, Mária Bieliková, Veronika Bogina, Tsvi Kuflik, and Roy Sasson (Eds.), Vol. 1922. CEUR-WS.org, 16–20. <http://ceur-ws.org/Vol-1922/paper4.pdf>
- [25] Bruce Ferwerda, Marko Tkalcić, and Markus Schedl. 2017. Personality Traits and Music Genres: What Do People Prefer to Listen To?. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, New York, NY, USA, 285–288. DOI: <http://dx.doi.org/10.1145/3079628.3079693>
- [26] Andrew J Flanagan and Miriam J Metzger. 2000. Perceptions of Internet information credibility. *Journalism & Mass Communication Quarterly* 77, 3 (2000), 515–540.
- [27] Norbert Fuhr, Mounia Lalmas, Saadia Malik, and Gabriella Kazai. 2006. *Advances in XML Information Retrieval and Evaluation: 4th International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- [28] John S Garofolo, Cedric GP Auzanne, and Ellen M Voorhees. 2000. The TREC Spoken Document Retrieval Track: A Success Story. *NIST SPECIAL PUBLICATION* 500, 246 (2000).
- [29] Alexandru L Ginsca, Adrian Popescu, and Mihai Lupu. 2015. Credibility in information retrieval. *Foundations and Trends in Information Retrieval* 9, 5 (2015), 355–475.
- [30] Jennifer Golbeck and Eric Norris. 2013. Personality, Movie Preferences, and Recommendations. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13)*. ACM, New York, NY, USA, 1414–1415. DOI: <http://dx.doi.org/10.1145/2492517.2492572>
- [31] John Hannon, Kevin McCarthy, James Lynch, and Barry Smyth. 2011. Personalized and automatic social summarization of events in video. In *Proceedings of the 16th international conference on Intelligent user interfaces*. 335–338.
- [32] Shaoyong Ji, Shirui Pan, Erik Cambria, Pekka Martinen, and Philip S Yu. 2020. A survey on knowledge graphs: Representation, acquisition and applications. *arXiv preprint arXiv:2002.00388* (2020).
- [33] Rosie Jones, Ben Carterette, Ann Clifton, Maria Eskevich, Gareth Jones, Jussi Karlgren, Aasish Pappu, Sravana Reddy, and Yongze Yu. 2020. Overview of the TREC 2020 Podcasts Track. In *The 29th Text Retrieval Conference (TREC 2020) notebook*. NIST.
- [34] Henry Kautz, Bart Selman, and Mehul Shah. 1997. Referral Web: Combining Social Networks and Collaborative Filtering. *Commun. ACM* (March 1997), 63–65.
- [35] Seyed Mehran Kazemi and David Poole. 2018. Simple embedding for link prediction in knowledge graphs. In *Advances in neural information processing systems*. 4284–4295.
- [36] Jin Young Kim and W. Bruce Croft. 2012. A Field Relevance Model for Structured Document Retrieval. In *Proceedings of the 34th European Conference on Advances in Information Retrieval (ECIR'12)*. Springer-Verlag, Barcelona, Spain, 97–108.
- [37] Irwin King, Michael R. Lyu, and Hao Ma. 2010. Introduction to Social Recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. 1355–1356.
- [38] Dominik Kowald, Markus Schedl, and Elisabeth Lex. 2020. The Unfairness of Popularity Bias in Music Recommendation: A Reproducibility Study. In *Advances in Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 35–42.
- [39] Mounia Lalmas and Anastasios Tombros. 2007. Evaluating XML Retrieval Effectiveness at INEX. *SIGIR Forum* 41, 1 (June 2007), 40–57.
- [40] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, and others. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
- [41] Lin-shan Lee, James Glass, Hung-yi Lee, and Chun-an Chan. 2015. Spoken content retrieval—beyond cascading speech recognition with text retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23, 9 (2015), 1389–1420.

- [42] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. 2010. Predicting Positive and Negative Links in Online Social Networks (*WWW '10*). 641–650.
- [43] Ang Li, Alice Wang, Zahra Nazari, Praveen Chandar, and Benjamin Carterette. 2020. Do podcasts and music compete with one another? Understanding users' audio streaming habits. In *Proceedings of The Web Conference 2020*. 1920–1931.
- [44] Junjie Li, Haoran Li, and Chengqing Zong. 2019. Towards personalized review summarization via user-aware sequence network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6690–6697.
- [45] Danyang Liu, Ting Bai, Jianxun Lian, Xin Zhao, Guangzhong Sun, Ji-Rong Wen, and Xing Xie. 2019. News Graph: An Enhanced Knowledge Graph for News Recommendation. In *KaRS@ CIKM*. 1–7.
- [46] Clifford A Lynch. 2001. When documents deceive: Trust and provenance as new factors for information retrieval in a tangled web. *Journal of the American Society for Information Science and Technology* 52, 1 (2001), 12–17.
- [47] Hao Ma, Tom Chao Zhou, Michael R. Lyu, and Irwin King. 2011. Improving Recommender Systems by Incorporating Social Contextual Information. *ACM Trans. Inf. Syst.* (April 2011).
- [48] Paolo Massa. 2013. A Survey of Trust Use and Modeling in Real Online Systems. (2013).
- [49] Nicolaas Matthijs and Filip Radlinski. 2011. Personalizing Web Search Using Long Term Browsing History. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11)*. ACM, New York, NY, USA, 25–34. DOI: <http://dx.doi.org/10.1145/1935826.1935840>
- [50] Kathleen McKeown, Julia Hirschberg, Michel Galley, and Sameer Maskey. 2005. From text to speech summarization. In *Proceedings (ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, Vol. 5. IEEE, v–997.
- [51] Matthew McLean. 2020. Podcast Discovery Stats in 2020: How Listeners Discover New Shows. *The Podcast Host* (Dec 2020). <https://www.thepodcasthost.com/promotion/podcast-discoverability/> Accessed Dec 2020.
- [52] Rishabh Mehrotra, James McInerney, Hugues Bouchard, Mounia Lalmas, and Fernando Diaz. 2018. Towards a Fair Marketplace: Counterfactual Evaluation of the Trade-off between Relevance, Fairness & Satisfaction in Recommendation Systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 2243–2251. DOI: <http://dx.doi.org/10.1145/3269206.3272027>
- [53] Alessandro B. Melchiorre and Markus Schedl. 2020. Personality Correlates of Music Audio Preferences for Modelling Music Listeners. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*. Association for Computing Machinery, New York, NY, USA, 313–317. DOI: <http://dx.doi.org/10.1145/3340631.3394874>
- [54] Stefano Mizzaro. 1997. Relevance: The whole history. *Journal of the American society for information science* 48, 9 (1997), 810–832.
- [55] Stefano Mizzaro. 1998. How many relevances in information retrieval? *Interacting with computers* 10, 3 (1998), 303–320.
- [56] Zahra Nazari, Christophe Charbuillet, Johan Pages, Martin Laurent, Denis Charrier, Briana Vecchione, and Ben Carterette. 2020. Recommending Podcasts for Cold-Start Users Based on Music Listening and Taste. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25–30, 2020*, Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu (Eds.), ACM, 1041–1050. DOI: <http://dx.doi.org/10.1145/3397271.3401101>
- [57] Nic Newman. 2019. Podcasts: Who, Why, What, and Where? <https://www.digitalnewsreport.org/survey/2019/podcasts-who-why-what-and-where/>. (2019). (Accessed on 02/09/2021).
- [58] Listen Notes. 2021. Podcast Stats. <https://www.listennotes.com/podcast-stats/>. (2021). (Accessed on 02/09/2021).
- [59] Jun Ogata and Masataka Goto. 2012. Podcastle: Collaborative training of language models on the basis of wisdom of crowds. In *Thirteenth Annual Conference of the International Speech Communication Association*.
- [60] Enrico Palumbo, Giuseppe Rizzo, and Raphaël Troncy. 2017. Entity2rec: Learning user-item relatedness from knowledge graphs for top-n item recommendation. In *Proceedings of the eleventh ACM conference on recommender systems*. 32–36.
- [61] Heiko Paulheim. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web* 8, 3 (2017), 489–508.
- [62] Pavel Pecina, Petra Hoffmannová, Gareth J. F. Jones, Ying Zhang, and Douglas W. Oard. 2008. Overview of the CLEF-2007 Cross-Language Speech Retrieval Track. In *Proc. CLEF*.
- [63] Maxime Peyrard. 2019. Studying Summarization Evaluation Metrics in the Appropriate Scoring Range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 5093–5100. DOI: <http://dx.doi.org/10.18653/v1/P19-1502>
- [64] PodBean. 2021. Free Podcast Hosting. available at <https://podbean.com> (accessed Feb. 9, 2021) (2021).
- [65] J. M. Ponte and W. B. Croft. 1998. A Language Modeling Approach to Information Retrieval. In *SIGIR '98*. 275–281.
- [66] PwC. 2020. Global Entertainment & Media Outlook 2020–2024. <https://www.pwc.com/outlook>. (2020). (Accessed on 02/09/2021).
- [67] Vahed Qazvinian, Emily Rosengren, Dragomir Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying misinformation in microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 1589–1599.
- [68] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting Extraneous Content in Podcasts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*.
- [69] Sravana Reddy, Yongze Yu, Aasish Pappu, Aswin Sivaraman, Rezvaneh Rezapour, and Rosie Jones. 2021. Detecting Extraneous Content in Podcasts. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 1166–1173. <https://www.aclweb.org/anthology/2021.eacl-main.99>
- [70] Peter Rentfrow, Lewis R. Goldberg, and Ran Zilca. 2011. Listening, Watching, and Reading: The Structure and Correlates of Entertainment Preferences. *Journal of Personality* 79 (4 2011), 223–258. DOI: <http://dx.doi.org/10.1111/j.1467-6494.2010.00662.x>
- [71] Peter J. Rentfrow and Samuel D. Gosling. 2003. The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology* 84, 6 (2003), 1236–1256.
- [72] Edison Research. 2019. The podcast consumer. available at <https://www.edisonresearch.com/the-podcast-consumer-2019/> (accessed Feb. 9, 2021) (2019).
- [73] Edison Research. 2020. The Infinite Dial 2020. <https://www.edisonresearch.com/the-infinite-dial-2020/>. (2020). (Accessed on 02/09/2021).
- [74] Pew Research. 10 facts about Americans and Facebook. available at <https://www.pewresearch.org/fact-tank/2019/05/16/facts-about-americans-and-facebook/> (accessed Feb. 9, 2021) (????).
- [75] S. Robertson and H. Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (April 2009), 333–389.
- [76] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 Extension to Multiple Weighted Fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM '04)*. ACM, Washington, D.C., USA, 42–49.
- [77] Tefko Saracevic. 1975. RELEVANCE: A review of and a framework for the thinking on the notion in information science. *J. Am. Soc. Inf. Sci.* 26, 6 (1975), 321–343. DOI: <http://dx.doi.org/10.1002/asi.4630260604>
- [78] Matthew Sharpe. 2020. A review of metadata fields associated with podcast RSS feeds. *arXiv preprint arXiv:2009.12298* (2020).
- [79] Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '96)*.
- [80] Mark D. Smucker and Charles L.A. Clarke. 2012. Time-based calibration of effectiveness measures. In *Proceedings of the 35th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR '12)*.
- [81] Harald Steck. 2011. Item Popularity and Recommendation Accuracy. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 125–132. DOI: <http://dx.doi.org/10.1145/2043932.2043957>
- [82] Y. Sun and J. Han. 2012. *Mining Heterogeneous Information Networks: Principles and Methodologies*.
- [83] Sean Szumlanski and Fernando Gomez. 2010. Automatically acquiring a semantic network of related concepts. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 19–28.
- [84] Jiliang Tang, Huiji Gao, and Huan Liu. 2012. MTrust: Discerning Multi-Faceted Trust in a Connected World (*WSDM '12*).
- [85] Jiliang Tang, Xia Hu, and Huan Liu. 2013. Social recommendation: a review. *Soc. Netw. Anal. Min.* (2013), 1113–1133.
- [86] Lei Tang and Huan Liu. 2010. *Community Detection and Mining in Social Media*.
- [87] Manos Tsagkias, Martha Larson, Wouter Weerkamp, and Maarten de Rijke. 2008. PodCred: A Framework for Analyzing Podcast Preference. In *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web (WICOW '08)*. Association for Computing Machinery, New York, NY, USA, 67–74. DOI: <http://dx.doi.org/10.1145/1458527.1458545>
- [88] Manos Tsagkias, Martha A. Larson, and Maarten de Rijke. 2010. Predicting podcast preference: An analysis framework and its application. *J. Assoc. Inf. Sci. Technol.* 61, 2 (2010), 374–391. DOI: <http://dx.doi.org/10.1002/asi.21259>
- [89] Gabriel Vigliani and Ichiro Fujinaga. 2016. Automatic Music Recommendation Systems: Do Demographic, Profiling, and Contextual Features Improve Their Performance?. In *Proceedings of the 17th International Society for Music Information Retrieval Conference, ISMIR 2016, New York City, United States, August 7-11, 2016*, Michael I. Mandel, Johanna Devaney, Douglas Turnbull, and George Tzanetakis (Eds.), 94–100. https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/044_Paper.pdf
- [90] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 417–426.
- [91] Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2019. Explainable reasoning over knowledge graphs for recommendation.

- In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 5329–5336.
- [92] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. TwitterRank: Finding Topic-Sensitive Influential Twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*. 261–270.
- [93] Zhou Xing, Marzieh Parandehgheibi, Fei Xiao, Nilesh Kulkarni, and Chris Pouliot. 2016. Content-based recommendation for podcast audio-items using natural language processing techniques. In *2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016*, James Joshi, George Karypis, Ling Liu, Xiaohua Hu, Ronay Ak, Yinglong Xia, Weijia Xu, Aki-Hiro Sato, Sudarsan Rachuri, Lyle H. Ungar, Philip S. Yu, Rama Govindaraju, and Toyotaro Suzumura (Eds.). IEEE Computer Society, 2378–2383. DOI: <http://dx.doi.org/10.1109/BigData.2016.7840872>
- [94] Longqi Yang, Michael Sobolev, Christina Tsangouri, and Deborah Estrin. 2018. Understanding user interactions with podcast recommendations delivered via voice. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, Sole Pera, Michael D. Ekstrand, Xavier Amatriain, and John O'Donovan (Eds.). ACM, 190–194. DOI: <http://dx.doi.org/10.1145/3240323.3240389>
- [95] Longqi Yang, Michael Sobolev, Yu Wang, Jenny Chen, Drew Dunne, Christina Tsangouri, Nicola Dell, Mor Naaman, and Deborah Estrin. 2019. How intention informed recommendations modulate choices: A field study of spoken word content. In *The World Wide Web Conference*. 2169–2180.
- [96] Longqi Yang, Yu Wang, Drew Dunne, Michael Sobolev, Mor Naaman, and Deborah Estrin. 2019. More than just words: Modeling non-textual characteristics of podcasts. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 276–284.
- [97] Dong Yu and Li Deng. 2014. *Automatic Speech Recognition: A Deep Learning Approach*. Springer Publishing Company, Incorporated.
- [98] Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. 2018. Neural Ranking Models with Multiple Document Fields. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 700–708. DOI: <http://dx.doi.org/10.1145/3159652.3159730>
- [99] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 2335–2344.
- [100] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.