Podcast Metadata and Content: Episode Relevance and Attractiveness in Ad Hoc Search

Ben Carterette[1], Rosie Jones[1], Gareth F. Jones[2], Maria Eskevich[3], Sravana Reddy[1], Ann Clifton[1], Yongze Yu[1], Jussi Karlgren[1], Ian Soboroff[4]
[1]Spotify, [2]Dublin City University, [3]CLARIN ERIC, [4]NIST
[1]United States and Sweden, [2]Ireland, [3]Netherlands, [4]United States

ABSTRACT

Rapidly growing online podcast archives contain diverse content on a wide range of topics. These archives form an important resource for entertainment and professional use, but their value can only be realized if users can rapidly and reliably locate content of interest. Search for relevant content can be based on metadata provided by content creators, but also on transcripts of the spoken content itself. Excavating relevant content from deep within these audio streams for diverse types of information needs requires varying the approach to systems prototyping. We describe a set of diverse podcast information needs and different approaches to assessing retrieved content for relevance. We use these information needs in an investigation of the utility and effectiveness of these information sources. Based on our analysis, we recommend approaches for indexing and retrieving podcast content for ad hoc search.

CCS CONCEPTS

• **Information systems → Test collections**; **Relevance assessment**; *Presentation of retrieval results.*

KEYWORDS

datasets, search, information retrieval, spoken content retrieval, podcasts, broadcast media

ACM Reference Format:
Ben Carterette[1], Rosie Jones[1], Gareth F. Jones[2], Maria Eskevich[3], Sravana Reddy[1], Ann Clifton[1], Yongze Yu[1], Jussi Karlgren[1], Ian Soboroff[4]. 2021. Podcast Metadata and Content: Episode Relevance and Attractiveness in Ad Hoc Search . 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, 5 pages. https://doi.org/10.1145/3404835.3463101

1 INTRODUCTION

Podcasts are a rapidly expanding as a popular medium for delivery of spoken audio content. As of 2021, more than 38 million podcast episodes are available online [22]. Given the amount of podcast material available, we believe it is increasingly important that it be fully searchable if it is to be fully exploited by users.

At present, most podcast search is done via *catalog match*, using show titles, episode titles, and sometimes metadata provided by podcast creators. This metadata is of highly varied quality and hence usefulness in supporting search operations. Recommendation from friends and family [16] remains in the top-three ways people find podcasts, while non-podcast-listeners in the same study say that they do not know how to find a podcast, or that they do not really know where to start.

In this paper we report experiments using the test collection from the TREC 2020 Podcasts Track which show that the automatically transcribed content of podcast episodes is more reliably useful for search than metadata provided by podcast creators.

2 PODCAST FORMAT

Podcasts are distributed as audio streams or files, through RSS feeds containing multiple metadata fields [2]. A podcast *show* has a title, description, language, consumption order (episodic or sequential), and a list of categories (e.g., News, Sports, Comedy) selected by the creator from a predefined taxonomy. A show typically has multiple *episodes*, which are the distinct audio files. Each episode has its own title, description, and other information. All this metadata may be noisy or inadequate [18].

Podcasting is typically a spoken-word medium. However, the relative ease and low cost of recording and publishing means there is great variability in the specifics. Podcast episodes have a wide range of lengths, from just a few minutes to hours, although they tend to be between half an hour and an hour long. Podcast content can vary in form, e.g. being scripted or informal dialogues. These can pose problems for search systems built for text archives, and also for the transcription of the content.

3 SPEECH RETRIEVAL

At first glance, podcasts are spoken documents, which have been well-studied in the TREC Spoken Document Retrieval Track which ran from 1997-2000 [9]. However, this work was based on news corpora, which are relatively homogenous in genre, style, and speaking professionalism, while podcasts come in many disparate forms, increasing the difficulty of the task. Retrieval from an archive of oral history has also been well-studied [15], but lacks the multi-speaker, multi-genre aspect of podcasts. The NTCIR Spoken Query and Document tasks [1] used a document collection of spontaneous speech, but the 600 hours of speech is tiny compared to the 50,000 hours of speech in the Spotify podcasts corpus [4]. The identification of “jump-in” points in multimedia content based on the spoken soundtrack at Mediaeval from 2011-2015 [6, 7, 14] motivates an approach to podcast search in which we identify the best place to start listening. We refer to Jones [10] for a more complete overview

SIGIR ’21, July 11–15, 2021, Virtual Event, Canada.

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463101

of research in spoken content retrieval from its beginnings in the early 1990s to today.

Because of the availability of metadata fields, podcasts can be represented as semi-structured documents, and models like BM25F [17], Field Relevance Models [12], and NRMF [23], can be adopted for podcast search tasks. Evaluation campaigns such as the INEX XML retrieval initiative [8, 13] have studied such models. As argued by Besser et al. [3], the goals of podcast search may be similar to those for blog search, if podcasts are viewed as audio blogs.

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. [20] 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.

4 TREC 2020 PODCASTS TRACK

The Podcasts Track ran for the first time at TREC 2020 [11]. It featured two tasks: a search task and a summarization task, the former is the focus of this paper.

Podcast search focused on a segment retrieval task defined as the problem of finding relevant segments of podcast episodes given a query representing a specific information need. The corpus for the task comprised 100,000 podcast episodes released in 2019, including metadata, audio, and full transcripts produced by automatic speech recognition (ASR) [4]. Participants were asked to retrieve unique episode identifiers (episode URIs) along with a time offset to the start of a two-minute segment starting on the minute within that episode. The segment corpus comprises 3.4M segments with an average word count of 340 ± 70 per segment.

Seven participating groups submitted a total of 24 runs to the search task. Each run ranked up to 1,000 segments for each of the 50 queries. Runs used retrieval techniques such as relevance feedback, query expansion, word2vec, BERT reranking, and fusion.

5 PODCAST SEARCH

We now take a step back to ask: why are we interested in this kind of podcast search? What are the user needs that can be addressed by retrieving segments of episodes? What other information could be retrieved that might improve the user experience?

5.1 Information needs for podcast search

User needs pertaining to podcasts include education, entertainment, and information. The format and variability of the podcast medium – series of episodes, the importance of the specific host or guest in the episode, range of presentation styles from monologues to interviews and banter, production quality – are additional aspects that may determine whether a user is interested in listening or not, entirely separate from topical relevance. These factors make podcast search different from traditional search tasks.

```
<topic>
<num>34</num>
<query>halloween stories and chat</query>
<type>topical</type>
<description>I love Halloween and I want to hear stories and conversations
            about things people have done to celebrate it.  I am not looking
            for information about the history of Halloween or generalities
                 about how it is celebrated, I want specific stories from
                 individuals.
</description>
</topic>
```

Figure 1: Example search topics

5.2 Topic development for TREC

Since there is currently no large-scale content-based podcast search engine, there is no simple source of sample topics. Thus we developed topics by introspection, considering what we might use a podcast search engine for, what we could imagine others using it for, what types of information needs make it more attractive than web search, etc. Topic development used several sources: lists of events in 2019, topic creators interests, and browsing metadata for potentially interesting content. To determine whether the topic would be interesting, we roughly compared metadata matches, web search results, and the results from our in-house transcript search engine. Topics were finally selected based on whether they retrieved interesting content from a simple search index of the podcast transcripts.

We also experimented with "known item" and "refinding" information needs. In total, eight development and 50 test topics were developed, with 13 topics within the test set being labeled as refinding or known-item.

5.3 Assessing relevance

5.3.1 TREC. Participant submissions were pooled to depth 20 (minimum 128 segments per topic, maximum 306) and reviewed by NIST assessors. The assessors primarily reviewed the ASR text of the segment, but could listen to audio if the need arose. The assessor's view showed each retrieved segment within the context of the transcript of the full episode, so that they could explore the context of the segment to better understand it. All segments from the same episode were judged in sequence together.

The relevance judgments were on a four-point scale of bad (0), fair (1), good (2), or excellent (3). For known-item and refinding topics, an additional level of "perfect" (4) was added and meant that the segment was precisely what the user was looking for. Excellent segments were completely on-topic, provided highly relevant information, and represented an ideal entry point into the episode.

5.3.2 Additional Assessments. The authors of this work independently assessed podcast metadata and full transcripts. From the metadata, we extracted episode titles and descriptions. Episode titles and descriptions are frequently too short or unspecified to accurately reflect episode relevance, so instead we assess how *attractive* they might be to a user with the stated information need.

We produced the following additional assessments:

- **Title attractiveness.** Would a user with the given information need find the episode a good candidate to stream based solely on the title?

source	no. topics	type	no. assessments
TREC	16	title	1,381
		title+description	1,381
		transcript-segment	1,863
		transcript-full	114
metadata +transcript indexes	50	title	1,049
		title+description	1,049
		transcript-segment	514
		transcript-full	254

Table 1: New assessments collected for this work. Assessments were collected from documents retrieved for a subsample of topics from TREC runs, and from all documents retrieved from indexes described in Section 6.

- **Title and description attractiveness.** Would the user find the episode a good candidate to stream based on the title and description together?
- **Full transcript relevance.** Is the episode relevant to the information need (based on the full transcript)?

We also used TREC segment-level judgments to obtain transcript-level judgments by taking the maximum relevance of any segment of an episode as the relevance of the transcript. We refer to this as the **transcript-segment** judgment. We acknowledge that these may be unreliable; it is possible that there is a segment of the episode more relevant than any seen by a TREC assessor and therefore our transcript-level judgment is low.

We selected documents to judge as follows: First, we built five new indexes using different combinations of metadata and transcripts; see Section 6 for details. We retrieved the top 10 results for all TREC topics from each of these five indexes, pooled them, and judged all titles and descriptions for attractiveness, plus a select subset of transcripts for relevance. This provided about 1,000 episode title and description attractiveness judgments, plus about 250 transcripts that had not previously been assessed for TREC.

We also pooled the top 10 results from all 24 TREC submitted runs. We selected a random sample of 16 topics for judging title and description attractiveness in this pool. We judged another 1,400 episode titles and descriptions in this set, as well as another 100 transcripts. Episodes were ordered by a function of rank position in runs, which may produce ordering effects [5]. We did not investigate this. Table 1 summarizes the new assessments.

There is very little overlap between episodes from our new indexes and episodes from TREC submitted runs. Across all 50 topics, there are only 79 episodes that occur in the top 10 of both our new runs and the 24 TREC submissions—only 7.7% of all episodes retrieved by our five new runs. By focusing on metadata we retrieve many more potentially relevant episodes.

6 INDEXING AND SEARCHING PODCASTS

An RSS feed with podcast episode titles, descriptions, and other metadata, in addition to the audio file, includes a lot of information that could be indexed for retrieval. In order to understand the relative utility of various free-text fields, we constructed Lucene indexes of episode titles, episode descriptions, titles and descriptions concatenated, full ASR transcripts, and transcripts concatenated with titles and descriptions. We retrieved rankings of episode URIs for all 50 TREC topics from each of these five indexes.

	assessment type			
index	title	title + description	transcript- segment	transcript- full
title	0.43	0.33	0.19	0.21
description	0.40	0.51	0.24	0.45
title+description	0.49	0.56	0.27	0.46
full transcript	0.37	0.42	0.41	0.52
transcript+title+description	0.43	0.51	0.45	0.61

Table 2: NDCG@10 results for five different indexes evaluated with four assessment types.

	assessment type			
run	title	title + description	transcript- segment	transcript- full
oudalab1	0.03	0.05	0.02	0.00
hltcoe1	0.17	0.23	0.18	0.26
hltcoe5	0.16	0.18	0.19	0.22
hltcoe3	0.15	0.19	0.24	0.20
hltcoe2	0.25	0.28	0.31	0.29
LRGREtvrs-r_3	0.20	0.29	0.32	0.37
BERT-DESC-TD	0.20	0.27	0.33	0.41
BM25	0.20	0.27	0.33	0.41
BERT-DESC-Q	0.20	0.27	0.33	0.41
RERANK-QUERY	0.20	0.27	0.33	0.41
RERANK-DESC	0.20	0.27	0.33	0.41
BERT-DESC-S	0.20	0.27	0.33	0.41
LRGREtvrs-r_2	0.24	0.30	0.35	0.36
QL	0.21	0.26	0.37	0.38
LRGREtvrs-r_1	0.25	0.32	0.37	0.38
hltcoe4	0.29	0.29	0.43	0.32
run_dcu1	0.33	0.37	0.44	0.44
run_dcu3	0.33	0.36	0.46	0.43
UTDThesis_Run1	0.28	0.32	0.46	0.28
run_dcu5	0.35	0.36	0.47	0.45
run_dcu2	0.33	0.37	0.48	0.44
UMD_IR_run2	0.32	0.34	0.48	0.37
run_dcu4	0.35	0.37	0.49	0.45
UMD_ID_run4	0.32	0.37	0.54	0.40
UMD_IR_run1	0.29	0.37	0.55	0.35
UMD_IR_run3	0.31	0.39	0.58	0.41
UMD_IR_run5	0.35	0.42	0.58	0.40

Table 3: Evaluation of TREC submitted runs and baselines by NDCG@10 with four new types of assessments.

We also obtained the TREC submitted runs. In order to make the segment rankings comparable to URI rankings from our indexes, we compressed all retrieved segments from one URI to a single result for that URI, ranked at the position of the top-ranked segment, and then removed all subsequent mentions of that URI.

Table 2 demonstrates how the choice of assessment and index impact retrieval results. Rows correspond to retrieval from the indexes described above. Columns correspond to the four different judgment types summarized in Table 1. Conclusions from this table:

(1) A ranking of episode titles from a title-only index is more attractive than a ranking of episode titles from a description-only index, while a ranking of descriptions from a description-only index is more attractive than a ranking of titles from

	transcript			description		% transcript words
	avg. length	vocab size	avg. ratio	avg. length	vocab size	in description
True Crime	2200	870	0.48	100	71	4.3
Religion and Spirituality	2000	700	0.44	79	55	3.5
Government	2500	860	0.41	82	60	3.1
Business and Technology	2900	910	0.38	141	94	4.1
News and Politics	2800	1200	0.46	100	72	2.8
History	2300	1000	0.49	81	58	2.9
Fiction	1900	770	0.52	78	53	2.6

Table 4: Terminological coverage of description with respect to transcript, content words

a description-only index, suggesting that titles and descriptions are not always signaling the same topical relatedness.

(2) The most attractive results are achieved by retrieving against an index of episode titles and descriptions.
(3) A ranking based on retrieval of transcripts returns many more relevant episodes, though titles and descriptions are much less attractive.
(4) Indexing transcripts and metadata together provides the strongest relevance by either transcript judgment type with a relatively small decrease in attractiveness.
(5) Result attractiveness remains surprisingly low in all cases—many results will appear not relevant.

From this we further conclude that there is a great deal of relevant material to be excavated from the transcripts that cannot be accessed via the metadata. But enabling users to find that information via full-document search means that result presentation will be negatively impacted. In other words, metadata search is inadequate for finding relevant information, but full-text search results in unappealing rankings. Thus some form of query-biased summarization [19] or passage retrieval is necessary for presenting results to users in an appealing way.

Table 3 shows the TREC runs evaluated by the four new assessment types. We observe the following:

- There is stronger correlation between different assessment types among TREC runs than among those in Table 2.
- By the transcript-segment judgments, our indexes are not competitive with the best TREC submitted runs.
- By the new transcript-level relevance assessments, all but one of our runs are better than the best TREC run.
- Our runs are substantially better than TREC runs on title and description attractiveness.

Based on these findings, it could be argued that indexing metadata and transcripts together, then searching full transcripts rather than segments, is likely to provide the best overall user experience.

7 ASSESSMENT AGREEMENT

Transcript relevance. Transcript-segment and transcript-full assessments agree that an episode is relevant in 71% of cases. 43% of episodes have agreement on the exact grade of relevance. This is a high level of agreement compared to other retrieval tasks [21].

Attractiveness and relevance. When either the title is judged attractive *or* the episode is judged relevant, in about 50% of cases both are true. This is a decent level of agreement for a retrieval task, but it means that episode title attractiveness is often not a very good predictor of episode relevance. The results are very similar for description attractiveness.

Interannotator agreement. URIs that overlapped between the five metadata runs and the TREC runs were independently judged by two different assessors. Among episode titles that at least one assessor marked attractive, assessors agreed 88% of the time. Similarly with descriptions, in 82% of cases both assessors found the description attractive. Agreement on exact grade of attractiveness was high as well: 72% for titles and 68% for descriptions.

8 TOPICAL FOCUS

Table 4 shows the coverage of content words for select categories of podcasts. Some genres have episodes with relatively high topical focus, which is reflected in the topical coverage of the descriptions. Each transcript and description is filtered to only contain nouns, verbs, and adjectives; the table reports counts of these tokens. The average ratio of vocabulary size to length, a measure of topical variation within a text, varies across the categories. While all these scores are normal, the topical variation of "Fiction" podcasts is considerably wider than that of "Business and Technology" podcasts. The last column of the table demonstrates a difference in terminological coverage. This can be understood as a measure of the topical representativity of the descriptions with respect to the episode. A higher score will mean that the description represents more of the topical content. The score variation indicates a potential for determining the topicality of the episode or show, and thus the utility of using search technology optimised for topical retrieval and content analysis (as opposed to usage-based similarity measures).

9 CONCLUSION

It is not difficult to find information needs for podcast search such that highly relevant content is buried in episodes, its presence not indicated by either episode title or description. Episode titles and descriptions that appear attractive may lead to irrelevant content and frustrated users. Podcast search engines should index both metadata and episode content, and episode segments or query-biased summaries may be necessary to help users understand why retrieved episodes are relevant to their need.

Acknowledgments Thanks to the reviewers for their constructive suggestions, and to TREC participants and organizers for helping make the track a success. Gareth Jones is partially supported by Science Foundation Ireland as part of the ADAPT Centre (Grant 13/RC/2106) at Dublin City University.

REFERENCES

[1] Tomoyosi Akiba, Hiromitsu Nishizaki, Hiroaki Nanjo, and Gareth J. F. Jones. 2011. Overview of the NTCIR-12 SpokenQuery & Doc-2 Task. In *Proc. NTCIR*.

[2] Apple. 2020. A Podcaster's Guide to RSS. https://help.apple.com/itc/podcasts_connect/#/itcb54353390. (Accessed on 01/30/2021).

[3] 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, Vol. 448)*.

[4] 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 *Proc. COLING'20*.

[5] Michael Eisenberg and Carol Barry. 1988. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *JASIST* 39, 5 (1988).

[6] Maria Eskevich, Robin Aly, Roeland Ordelman, David N. Racca, Shu Chen, and Gareth J. F. Jones. 2015. SAVA at MediaEval 2015: Search and Anchoring in Video Archives. In *Working Notes Proceedings of the MediaEval 2015 Multimedia Benchmark Workshop*.

[7] Maria Eskevich, Gareth J. F. Jones, Shu Chen, Robin Aly, Roeland Ordelman, and Martha Larson. 2012. Search and Hyperlinking Task at MediaEval 2012. In *Working Notes Proceedings of the MediaEval 2012 Multimedia Benchmark Workshop*.

[8] 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*.

[9] John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. 2000. The TREC Spoken Document Retrieval Track: A Success Story (RIAO). In *Content-Based Multimedia Information Access - Volume 1*.

[10] Gareth J. F. Jones. 2019. About Sound and Vision: CLEF Beyond Text Retrieval Tasks. In *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer.

[11] 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 *Proc. TREC*.

[12] Jin Young Kim and W. Bruce Croft. 2012. A Field Relevance Model for Structured Document Retrieval. In *Proc. ECIR*.

[13] Mounia Lalmas and Anastasios Tombros. 2007. Evaluating XML Retrieval Effectiveness at INEX. *SIGIR Forum* 41, 1 (June 2007), 40–57.

[14] Martha Larson, Maria Eskevich, Roeland Ordelman, Christoph Kofler, Sebastian Schmiedeke, and Gareth J. F. Jones. 2011. Overview of MediaEval 2011 Rich Speech Retrieval Task and Genre Tagging Task. In *Proc. MediaEval*.

[15] 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*.

[16] Edison Research. 2019. The podcast consumer. *available at https://www.edisonresearch.com/the-podcast-consumer-2019/ (accessed Feb. 9, 2021)* (2019).

[17] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 Extension to Multiple Weighted Fields. In *Proc. CIKM*.

[18] Matthew Sharpe. 2020. A review of metadata fields associated with podcast RSS feeds. In *PodRecs: Workshop on Podcast Recommendations*.

[19] Damiano Spina, Johanne R. Trippas, Lawrence Cavedon, and Mark Sanderson. 2017. Extracting audio summaries to support effective spoken document search. *Journal for the Association of Information Science and Technology (JASIST)* 68, 9 (2017).

[20] 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* (Napa Valley, California, USA) *(WICOW '08)*.

[21] Ellen M. Voorhees. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proc. SIGIR*.

[22] Gavin Whitner. 2021. The Meteoric Rise of Podcasting. https://musicoomph.com/podcast-statistics https://musicoomph.com/podcast-statistics (Accessed February 2021).

[23] Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. 2018. Neural Ranking Models with Multiple Document Fields. In *Proc. WSDM*.