On Natural Language User Profiles for Transparent and Scrutable Recommendation

Filip Radlinski
Google
London, UK
filiprad@google.com

Krisztian Balog
Google
Stavanger, Norway
krisztianb@google.com

Fernando Diaz
Google
Montreal, Canada
diazfernando@google.com

Lucas Dixon
Google
Paris, France
ldixon@google.com

Ben Wedin
Google
Cambridge, USA
wedin@google.com

ABSTRACT

Natural interaction with recommendation and personalized search systems has received tremendous attention in recent years. We focus on the challenge of supporting people's understanding and control of these systems and explore a fundamentally new way of thinking about representation of knowledge in recommendation and personalization systems. Specifically, we argue that it may be both desirable and possible for algorithms that use natural language representations of users' preferences to be developed. We make the case that this could provide significantly greater transparency, as well as affordances for practical actionable interrogation of, and control over, recommendations. Moreover, we argue that such an approach, if successfully applied, may enable a major step towards systems that rely less on noisy implicit observations while increasing portability of knowledge of one's interests.

CCS CONCEPTS

• **Information systems** → **Personalization**; **Recommender systems**.

KEYWORDS

recommendation; transparency; scrutability; natural language

ACM Reference Format:
Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On Natural Language User Profiles for Transparent and Scrutable Recommendation. In *Proceedings of the 45th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3477495.3531873

1 INTRODUCTION

Personalized search and recommendation systems rely on a user representation to evaluate the match between a person and candidate results. These representations are often difficult to interpret, modify, and explain [89]. Whether the information need is a standing interest (such as with movie recommendation) or acute (such as when someone has a current information need and issues a search query), personalized systems rely on some representation to identify which results have maximal value. For example, a common representation of a user is an embedding in a continuous vector space [67]. As such representations can be complex, systems rarely show them to the person receiving recommendations. Moreover, when representations are difficult for people to interpret, even when a user can see them they are unlikely to help the user understand or control the recommendations they receive.

We contrast this common approach with a proposed future where systems represent users explicitly with *scrutable* natural language (NL); where scrutable language is defined as being both short and clear enough for someone to review and edit directly. Ideally, these natural language statements of preferences correspond to how users would either *choose to* describe their own interests in a given context, or *agree with* were this description presented to them. We envision such representations providing transparency, and allowing control of a system's personalization. For example, such NL user models could support people updating their representation when their preferences change, distinguishing between what they would like in the future from how they behaved in the past [26]. Traditional systems can be slow to recognize and adapt to such changes. Moreover, should a person choose to maintain a summary of their interests elsewhere, they could choose when and how to share it with other services they interact with. In terms of recommender quality, recent research suggests that even a simplistic natural language summary of a user's interests can provide recommendations not far from those generated by collaborative filtering methods [5].

We present a decomposition of the difficult challenge posed by natural language user representations into two steps: (1) obtaining a reliable, complete, yet succinct natural language description of a user's interests, and (2) recommending items based (exclusively) on this description. The first task, ***scrutable user model generation***, involves taking input similar to that used by systems today (be these explicit user preferences, ratings over sets of items, or past interactions) to generate a length-bounded natural language summary that may be scrutinized by the user. Most directly related to explainable recommendation [89], the natural language user model *describes what constitutes relevant results* and can also be utilized to generate *genuine explanations for specific item recommendations*. The second

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3531873

task, ***scrutable NL-based recommendation***, then takes a natural language user model as input to assess the match between items and the user model. While most related to long-form text retrieval [31], this task is significantly more challenging in that the long form must represent the person's interests holistically, rather than being focused on one acute need. In other words, long-form retrieval most often presents a set of conditions that must all be satisfied, while in the recommendation setting the goal is to capture the breadth of the recipient's interests (often along with context). We also note that traditionally recommendation and personalization place no constraints on the intermediate representation, and optimize user model generation and recommendation jointly. We argue that there are transparency and scrutability benefits with a NL constraint. Given the current state of the art, it may also be valuable to consider these tasks independently, drawing on advances in large scale models for natural language understanding to accelerate progress.

There are, of course, many challenges that would need to be addressed to reach state of the art performance in personalization and recommendation while obtaining the benefits we present. Our primary goal is to expand upon these challenges, collating evidence from recent work suggesting the feasibility of these being solved, and ultimately motivate progress in addressing them. While we do not expect this to yield immediate solutions, we argue they are worth investigating due to the potential larger value to society.

In summary, our main contributions are (1) a formal proposal of a fundamentally new approach to recommendation, where systems solely rely on a natural language user model that is fully transparent and controllable by the user; (2) a tractable decomposition of the approach, collating recent advances that may enable significant progress by the community; (3) a detailed presentation of specific open research challenges; (4) a qualitative feasibility study showing how recent advances in language models provide evidence towards the tractability of the problem.

2 OVERVIEW

We argue for a new philosophy for recommendation and personalization, one based on succinct natural language summaries of users' preferences rather than mathematically convenient numerical representations. These summaries may be inferred from implicit behavior data, explicit ratings and reviews, or provided directly by the user. The key principle is for people to more easily scrutinize representations of their interests—to understand what is happening, correct the system, and enable capturing aspirations not reflected in past behavior without significant reduction in recommendation quality. The algorithmic requirement enabling this is that personalization is made solely[1] on the basis of natural language summaries.

2.1 Main Benefits over Existing Approaches

We now review the main benefits of the proposed approach over contemporary recommender and personalization systems, which primarily rely on user-item ratings or assumptions about implicit interaction-based indicators of user-item match [67].

The most obvious benefit is *transparency*: by representing knowledge about users in natural language, it can be clearer to recipients

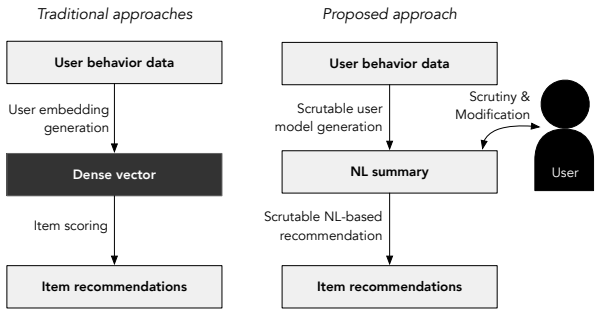

Figure 1: High-level overview of traditional approaches vs. our proposal. Instead of relying on latent representations of user preferences, we propose natural language as an explicit, transparent, and editable intermediate representation.

(i.e., users) what information is being used to generate recommendations. While explainable recommendation is a well recognized research area [89], most work has focused on explaining *item recommendations*, and often only captures a simplified summary of algorithm behavior as *post hoc* justifications. We focus on *user-level summaries*, as first suggested in [5], to suggest treating the user model in an easily accessible form: natural language. Such summaries would ideally make clear the aspects of a person's interests involved in the recommendation process. If successful, it would make explicit—and thus enable the elimination of—any undesirable criteria. At the same time, natural language summaries could support highly personalized *item-level explanations*, grounded in specific user model statements (see, e.g., Table 3).

A second benefit with more direct impact on recommendation quality is that such an approach can permit user summaries to be *scrutinized* (i.e., inspected and modified), most obviously to correct errors. Expanding significantly upon Balog et al. [5], we argue that it would be valuable for recommender systems to allow *any free-form changes*. This may provide particular value for systems that rely on implicitly observed user interactions. For instance, it has been observed that a temporary interest in a topic can lead to recommender systems presenting that topic over an extended period [52]. If a person could easily scrutinize their model, they could eliminate such incorrect inferences. While keyword-based introspection has been used for control of recommendations (e.g., [15]) and advertisements (e.g., [61]), in practice the volume of keywords at different levels of granularity can lead to complex user interfaces that require a high amount of effort to modify downstream outputs. With this in mind, we argue that in order for the scrutability benefit to be fully realized, the effort required to inspect and modify user representations to achieve the desired result downstream must be minimal in order to motivate meaningful user adoption.

If editing preference representations improves users' recommendations, then an additional systemic benefit emerges: there is an increased incentive for people to express their preferences. This in turn can improve future recommendations. Providing a feedback mechanism on aspects of the NL user summary could invite efficient user feedback about the user model—both about aspects that are

[1]The value, as well as the limitations introduced by this constraint are discussed directly in Sec. 5.3.3.

Table 1: Illustrative examples of traditional rating-based observations being summarized in natural language, scrutinized by a user, then transformed into a more informative descriptions yielding better recommendations in two domains.

Recommendation Stage	Domain	
	Movies	Events
Item Ratings	Shawshank Redemption (+), Persona (+), Star Trek (+) The Piano (+), Legend of 1900 (+), Kill Bill (+) The Hangover (-), Saving Private Ryan (-), Memento (-)	Cirque du Soleil (+), The Lion King musical (+) Monster Truck (+), Cambridge Beer Festival (+) Arsenal vs Chelsea football (-), Stomp (-)
Initial Generated NL Summary	I like dramas involving unique personalities, science fiction, and movies about prison. I don't like romantic comedies or movies about World War II. I don't like complicated movies.	I like musicals and acrobatics. I also like beer events and stunt driving. I don't like football events nor music events.
Initial Recommendations	Mission to Mars, The Waterboy, The Green Mile, Felon	Wicked (musical), Traveling circus, Heineken brewery tour, Wimbledon tennis
User-Scrutinized NL Summary	I like dramas involving unique personalities. I don't like movies about drunken behaviour nor young people's antics, but do like more refined romantic comedies. I don't generally like movies with a lot of violence, unless they have great cinematography like Quentin Tarantino. I also like positive science fiction stories, tho not dystopian ones, especially at the end of the week. I also like Pixar movies but have seen them all.	I like musicals and acrobatics. I make my own beer, so enjoy microbrewery events and tours. I like motor sports, racing and stunt driving, although I don't like any other sports events. I like going to outdoor markets on sunny days, but wouldn't travel far for one.
Scrutinized NL Recommendations	Roman Holiday, No Country for Old Men, Arrival, The Iron Giant	Billy Elliot: The Musical, Saturday SoHo Farmer's Market, Brewery Tour with Mikkeller

correct and incorrect. This novel type of feedback has the potential to create yet more value for recipients of recommendations.

A related benefit of direct user control is the ability to handle *significant shifts* in user preferences. It has been noted that people may find their interests change, for instance due to changing relationships, context, and so forth [33, 62]. In traditional recommendation or personalization algorithms, there are often few effective ways for someone to maintain recommendation quality while altering preferences significantly. Given a succinct user model, we hypothesize that it would be possible to modify just enough to cause the desired shift in recommendation behavior. An additional advantage of natural language user models is that they allow for explicit expressions of aspirations (that is, allowing a distinction between what someone would like to have recommended in the future versus how they behaved in the past) [26].

Finally, the existence of a user-accessible summary could also improve *cold-start recommendations*: the person may choose to describe their interests, bootstraping their recommendations. As discussed in more detail later, it may also be more useful for people to write a new, or copy an existing NL user summary.

2.2 Research Context

It is important to consider the question of what has changed recently to allow this fundamentally different way of representing knowledge for recommendation.

Foremost, the ability of machine learning systems to interpret, generate, and reason about natural language text has advanced tremendously, and has manifested in systems like GPT-3 [17] and LaMDA [76]. Much smaller models trained to look up entities have also shown promise on traditional recommendation tasks [87]. This progress suggests the plausibility of recommendation systems that use natural language as the primary knowledge representation.

At the same time, transparency and user control have received significant recent interest from the research community [56, 89]. Natural language representations of user interests offer the potential for a step change in how this can come about.

Finally, we observe that over the last few years new experimental approaches have led to datasets that, for the first time, capture both natural language and item-based knowledge about users, allowing the approach that we advocate to be evaluated directly [4, 5, 14, 63].

2.3 Main Components

The challenge we describe can naturally be decomposed into two components, illustrated in Figure 1, corresponding to (1) *scrutable user model generation*, and (2) *scrutable NL-based recommendation*. We illustrate these tasks with running examples in Table 1, showing how someone could be provided with a NL summary of their interests, modify it, and subsequently have a better experience.[2]

Definition 2.1 (Scrutable user model generation)**.** Given a set of interactions between a user and items (e.g., ratings, reviews, view-time, etc.) as well as metadata about items (e.g., product descriptions), *scrutable user model generation* refers to the task of generating a natural language summary (*NL summary*) within a given length bound that accurately summarizes a user's preferences in support of future recommendations and user scrutiny.

It is important to note that succinctness (ensured by having a maximum length) is a necessary condition of the representation. Intuitively, the longer the NL summary, the more information about a user's preferences can be accurately captured, but the increased length also increases the cognitive load needed to scrutinize—this trade-off is one of the key challenges (see Sec. 5.4.1). A key underlying assumption behind our approach is that the NL summary is short enough for the user to meaningfully scrutinize it, in particular to understand and modify their representation. Similarly, the scrutinization process must be easy enough, and result in visible improvements for users quickly enough, to motivate them to put in such effort.

We note that scrutable user model generation depends on the existence of text that can characterize items and their attributes. For example, this could be learned by a model from a knowledge base, associated reviews, explicit feedback given by users, etc. The source of such text is purposefully not made part of the definition, although we elaborate on various sources in Sec. 5.1.1.

[2]While Table 1 considers two popular recommendation domains, we argue that such an approach can be appropriate for almost any domain where user interests or needs can be summarized in natural language.

NL summaries may be evaluated *intrinsically*, in terms of the quality of the generated text, either overall or along specific dimensions (e.g., fluency, coherence, correctness) [19]. Ultimately, NL summaries also need *extrinsic* evaluation, measuring their utility for the end-to-end task of recommending useful items. Additionally, *scrutability* must clearly be a criterion: To what extent does the summary enable someone to understand their recommendations, and to what extent does it allow them to correct or update it?

Definition 2.2 (Scrutable NL-based recommendation)**.** Given a NL summary representing a user's preferences, *scrutable NL-based recommendation* refers to the task of generating a partial ordering over a set of candidate items that can be recommended.

Typically the ordering will be a top-N ranked list of item recommendations or a scoring of the items, based on predicted utility. For instance, see the "Recommendation" rows in Table 1 that could be generated from the preceding row's NL summary. This task can naturally be evaluated in terms of effectiveness as a top-N ranking task, using standard retrieval measures (see Sec. 4.2.1).

Joint development. The two components, generation and recommendation, are related to well-studied problems, as discussed in Sec. 3 and 4. However, our proposal is unique in that generation and recommendation jointly depend on the textual representation. Generation must produce text that is both interpretable and modifiable, and effective in producing good recommendations, and for its part the NL-based recommender can only be as good as the NL summaries it is provided with.

3 SCRUTABLE USER MODEL GENERATION

We start by considering in more detail the question of generating a natural language user summary, such as those illustrated in Table 1. Specifically, we propose the research problem of the automated generation, and updating of, scrutable natural language user summaries. For a given user, the output is a proposed new, or updated, natural language representation that they can scrutinize.

3.1 Research Context

Producing a description that is correct yet easy to interpret, captures someone's breadth of interests, and can be leveraged by a recommender to produce near state of the art recommendations is an open problem. To train a model that produces such a representation one would expect to require significant training data, which may appear difficult to obtain. However, we note three branches of work that give reason to be optimistic, and we argue that these may productively be brought together to tackle the open problem.

3.1.1 Fluent Generation: From Templates to Large Language Models. Balog et al. [5] recently proposed the notion of transparent, scrutable, and explainable user models for recommendation. Although they used a restricted representation of users as weighted pairs of tag interactions, the authors showed how representations can be verbalized as NL statements using templates (e.g., "You don't like [tag1] unless [tag2], for example [movie]"). Editing was limited to removing parts of these statements, which allowed the recommendation model to still use only (pairwise) tags.[3]

A key difference with the vision we lay out is that we are proposing that the generation of user summaries could be powered by advances in natural language generation rather than being template-based. This, we argue, can obtain much more nuanced statements of preference, allowing much more precise characterization of user interests. We will consider how these can be used for recommendation in Sec. 4. In particular, we note that recent large pre-trained language models have been shown to have impressive generative fluency [66, 76] as well as effective few-shot learning capabilities [17]. There has also been a recent branch of work that shows that a tiny fraction (e.g., 0.01%) of the model's parameters can be tuned to provide comparable quality to full fine-tuning [47], which suggests that bridging between vector-based models and textual ones for user representations may be feasible. We also note that while language model advances have often focused on chat settings, recent work has also illustrated recommendation use cases [76].

3.1.2 Effective User Characterization. It has long been recognized that tags can be used effectively for recommendation [13, 25, 27, 41]. They can also be used for explaining recommendations—as particularly well studied in the movie domain [30, 81]—as well as for tasks such as assisting user decision making [2]. Early work on preference learning suggests that it is easier for users to express preferences over item sets than on individual items [16, 23]. Indeed, Chang et al. [21] demonstrate that new users can complete preference elicitation more effectively by expressing preferences for groups of items (tags) rather than rating individual items. Given that tags are a shared expression of social communities, the tags users self-assign are highly influenced by the community [70, 84]. Sen et al. [69] argue that systems should also support negative user ratings on tags, not only positive ones. By using NL summaries, users are not limited to existing tag vocabularies and can also express the intensity of their preferences (e.g., 'don't like' vs. 'hate').

3.1.3 User Control. Past research shows that providing users with more control leads to higher satisfaction [15, 24, 36, 42, 60]. Jannach et al. [39] define user control mechanisms as requiring to have an *immediate effect* on the recommendations. In many real-world recommender systems, users have no other means to control the behavior of the system than to change their past ratings—this does not count as a control mechanism "as usually these changes are not immediately reflected in the results" [39]. User control may be exercised during the preference elicitation phase, where users can explicitly declare their preferences on certain content categories or assign weight to item attributes [38, 39]. Another form of control is to dynamically adjust the recommendations once they have been presented, using various UI elements for fine-tuning results [15, 24, 36, 40]. Harambam et al. [32] experiment with distinct control mechanisms for different phases in the recommendation process: input, recommender algorithms, and output. While having control empowers users, having too many controls also increases cognitive load [40] and the design of intuitive yet easy-to-use interfaces remains an open challenge and an active research area [39, 49].

[3]It should also be noted that the recommender algorithm uses tag weights from the underlying user model, relying on information not contained in the textual summary.

Our proposal provides control through scrutable NL summaries, allowing users to make updates and immediately see the results.

3.2 Evaluation and Datasets

3.2.1 Evaluating NL Summaries. Scrutable user representations play two key but very different roles: on the one hand, they support good recommendations; on the other, they should be scrutable, supporting the user's understanding and control.

A key evaluation of the *effectiveness* of generated NL summaries is *extrinsic*, that is, via the quality of the recommendations it allows. Because the generator's input includes interactions (e.g., ratings), recommendation quality is likely to depend on the number of past interactions. We thus propose to evaluate effectiveness as a function of past interactions, which also allows for a direct comparison with contemporary (rating-based) recommender systems. It is expected that, while the performance of contemporary recommender systems typically increases with the number of past interactions, it may be increasingly challenging to effectively capture additional information in a NL summary. This, however, remains an open question that can only be answered empirically.

Scrutability of a NL summary involves assessing its ability to be both understandable and editable. A better ground truth than asking users to write preferences may be for them to select or edit pre-written phrases: while domain experts (e.g., movie critics) can fluently express nuanced characterizations, users may struggle with precise vocabulary to describe their tastes beyond general terms, e.g., "I like sci-fi" [63]. It may be possible for such phrases to be extracted from reviews (e.g., by methods like [53]), be written by experts, or even be directly generated by specialized language models (cf. Sec. 6). This turns the significant challenge of writing preferences into one of selecting and validating them. Editability aims to provide users with control over their recommendations, by altering their profiles (i.e., going from Row 2 to Row 4 in Table 1). That said, the ambiguity of language is also important to consider [64], to ensure that the users and the system interpret the meaning of language consistently. Otherwise, a user's expression of a preference may lead to poor recommendations—suggesting that correctness of interpretation must also be evaluated as part of scrutability.

Scrutability can be evaluated quantitatively, by measuring how effective an edit was in achieving the desired outcome—e.g., by removing some specific unwanted items from the suggestions, or improving the quality of recommendations as a whole, as in [5]. The usability and effectiveness of the user interface also plays an important role here, and would require an appropriate evaluation. This is key, as both the effort involved and how quickly users see a result from scrutinizing their summary are likely to strongly influence whether or not users actually scrutinize the model, even if it is possible through natural language.

3.2.2 Evaluating Explanations. While receiving good recommendations is key, transparency and trustworthiness are also key user considerations [55, 71]. This is where explanations can help users to understand why a given item was suggested [78]. The NL summary establishes a shared language between the user and the system, which in turn allows for a greater degree of personalization to the individual. Item-level explanations have been evaluated in the past along different dimensions, such as transparency, scrutability, or effectiveness [78], as well as on how personalized they feel [4, 77]—we argue that similar evaluation criteria should be adapted for user-level summaries.

3.2.3 Existing Datasets. Generating high quality datasets of natural language representations is challenging. However, some existing datasets already highlight preliminary evaluation possibilities.

Radlinski et al. [63] used a Wizard-of-Oz approach to mimic a digital assistant that elicits movie preferences. While this study does not yield NL summaries, it provides useful insights into how users describe preferences in the movie domain. Separately, Balog et al. [5] generate template-based NL user summaries from item ratings, with users asked to scrutinize their summaries by inspecting individual sentences, and selecting a variant that most accurately described their preferences. Then, the corrected summaries are compared against initial ones in terms of recommendation quality. Narrative-driven recommendation datasets could also be used, trying to infer summaries from recommendations that follow [1, 14].

3.2.4 Limitations and practical considerations. Ben-Akiva et al. [8] note that stated preferences (what people say they prefer) may be inconsistent with their revealed preferences (what they actually select). This highlights a challenge with the proposed approach: people may act inconsistently. The effect is twofold: on the one hand, NL summaries that accurately reflect user behavior may not match what a user would say they like. Conversely, if the NL summary does accurately reflect a user's expressed interest, what they choose to consume may not match their initial expressed preferences. A related problem is that by leveraging language models, additional concerns about the unintended biases or other possible harms of utilizing such systems may need investigation [9, 91]. For instance, when characterizations are associated with specific items, it would be important to ensure that those characterizations are correct.

4 SCRUTABLE NL-BASED RECOMMENDATION

The second component, *scrutable NL-based recommendation*, is concerned with generating novel item recommendations based on a natural language user preference summary, reflected in moving from the second to third and fourth to fifth rows of Table 1. It involves a deep understanding of natural language and how to interpret it. We discuss similarities and key differences to related work, then present evaluation objectives and related datasets.

4.1 Research Context

Although natural language is a novel way to encode long-term user preferences, matching text against items has many connections to existing work. We review these similarities and contrast scrutable NL-based recommendation.

4.1.1 Filtering. Originally proposed by Luhn [48], information filtering refers to detecting documents relevant to an information need. In early systems, user profiles were constructed by selecting keywords in context, which were then matched against incoming documents [37]. Some years later, the Text REtrieval Conference (TREC) included a document routing task, bringing a common evaluation protocol, data, and metrics to information filtering [83].

Filtering profiles included keyword-like representations as well as longer sentence or paragraph-length profiles about user needs [35]. Underlining the similarity between filtering and recommendation, these topics were also used for TREC *ad hoc* search tasks. Here longer, user-composed 'narrative' descriptions included a variety of language, including restrictions, negations, and other modifiers [35]. When provided the opportunity to amend profiles, early information filtering system users were not able to effectively improve performance [37]. More generally, later studies have shown that personalization systems that rely on content-based models can improve search results [51, 75].

Although work in profile-based information filtering has similarities to scrutable NL-based recommendation, there are substantial differences. To start, we do not necessarily constrain our work to text documents or even streams of items. This means that, while early filtering systems may be similar in spirit, the precise techniques will be quite different from those specific to text. More importantly, we believe that language technology has sufficiently advanced that even early filtering approaches deserve reevaluation. This is especially true of previous results around user-editing of profiles, which may have suffered from keyword-based profiles.

4.1.2 Verbose Query Retrieval. Given the similarity of recommendation and search [7, 14, 28, 86] and the interchangeable nature of filtering profiles and queries, we next examine relevant work in information retrieval. Retrieval with verbose natural language queries has attracted much attention [31], yet the definition of a verbose query differs across retrieval contexts. In web search, queries with five words are considered long [31]. Narrative queries in the TREC Robust track were about 40 words long [82]. In clinical retrieval, queries as long as 60 words occur [43], while in community QA the median question length may be over 150 words [79].

There are studies showing that verbose queries perform better than short queries, e.g., [88]. However, in an analysis of hundreds of queries in TREC datasets, Bendersky and Croft [10] find that short queries, using the 'title' of the TREC topic definition, consistently outperform long 'description' queries. Another study on web search query logs shows that search engine performance drops significantly with increased query length [11]. Other techniques developed specifically for verbose queries include query reduction (e.g., deleting terms [85] or selecting key noun phases [10]) and assigning different weights to query terms based on their estimated importance [12, 57, 92]. In general the brittleness of retrieval systems to verbose queries may be due to the presence of more complex search strategies such as negation and hedging [3]. Specifically, while verbose queries, such as found in TREC, can include 'negative' preferences, retrieval performance can drop substantially when these expressions are included for standard retrieval engines [43].

While verbose query retrieval is related to scrutable NL-based recommendation, there are substantial differences. On the one hand, as with filtering, verbose query retrieval emphasizes text document retrieval, while our framing is much more general, including a variety of media. More specifically, information retrieval usually focuses on acute, focused information needs. Acute or ephemeral information needs contrast with the more persistent standing preferences found in recommendation. Focused information needs are expressed as conjunctions of concepts or facets with a small set of relevant documents. In our setting, a natural language profile covers a broad disjunction of preferences with a set of overlapping clusters of relevant items. This property of user profiles captures the importance of diversity and serendipity in recommendation [22, 58].

4.1.3 Narrative-driven Recommendations. The task of *narrative-driven recommendations* (NDR), introduced by Bogers and Koolen [14], describes a "scenario where the recommendation process is driven by both a log of the user's past transactions as well as a narrative description of their current needs or interests." NDR may be regarded as a specific form of context-aware recommendations, where recommendations not only correspond to a user's (implicit) preference profile ("items I would like"), but are also tailored to a given situation or context. NDR could also be seen as a personalized search task, given that the user is explicitly soliciting suggestions.

Common to NDR and scrutable NL-based recommendation is the recognition of the power of natural language in describing preferences and the context for the desired recommendations. A key difference, however, is that NDR maintains transaction logs as one of the main data sources (in addition to narrative descriptions). In scrutable NL-based recommendation we rely solely on the NL summary as input to the recommendations. Narrative-driven recommendations are most commonly solicited on community QA fora [1, 14], where there is often back-and-forth between the requester and those who give recommendations. We envisage that such user feedback could be elicited using conversational interfaces and in turn be utilized to update the NL summary (cf. Sec. 5.2).

Similarly, our proposal contrasts with *conversational recommendation* (e.g., [74]), where the user's interests may be solicited in a natural language back and forth. However, such systems tend not to summarize the need in natural language in a way that can be scrutinized and freely updated before results are selected for users.

4.1.4 Large Language Models. Recent advances in language modeling provide qualitative examples that suggest large language models can recommend items based on short utterances and dialog [76]. Our work proposes a significant extension of this hypothesis: that such models could allow NL summaries to become a shared scrutable object for both people and recommenders.

Sachdeva and McAuley [68] provide a review of approaches leveraging review text to develop user summary highlights and conclude that the progress to date has been relatively small. The position we take, however, is *not* that recommendations would necessarily need to be significantly *better*, but that a NL-based user model may offer both *good* recommendations, and *scrutable* ones, allowing users to inspect and edit their representation.

The closest related work that uses natural language to generate recommendations is that of Zemlyanskiy et al. [87]. They use a BERT-style language models to generate movie recommendations from free-form descriptive text. The main differences to our case is that these descriptions should holistically capture the users' broad interests. Narang et al. [53] demonstrate that large language models can also successfully extract relevant sub-parts of text (e.g., sentiment of movie reviews). Although they did not apply their models to extracting preferences, doing so would provide a significantly cleaner dataset for approaches like [87]. Moreover, this work combined with recent parameter efficient training methods, like

that of Lester et al. [47], suggest that significant further quality improvements are attainable.

4.2 Evaluation and Datasets

We define scrutable NL-based recommendation as a ranking task, with a system presenting items in decreasing order of predicted utility. In that sense, we can evaluate the effectiveness of a system ranking using standard information retrieval metrics such as reciprocal rank and nDCG. Yet we now explore how moving from relevance to utility requires a shift in evaluation methodology.

4.2.1 Evaluation Criteria. Modern systems commonly model how users interact with system rankings, often adopting retrieval metrics (such as precision or nDCG) [80]. Although an item rating, like relevance, encodes the affinity of a user to an item, when seeking recommendations, users are usually explicitly looking for items they have *not* previously experienced [29, 46]. This means that, unlike in search, the utility from an item depends on whether or how frequently an item has been experienced by the user. Additionally, the diversity of the recommendations may also be considered [18].

4.2.2 Existing Datasets. Several existing datasets can be leveraged for evaluating scrutable NL-based recommendation. As mentioned above, TREC filtering and *ad hoc* retrieval datasets include verbose narratives that contain challenging language, including restrictions and difficult to support negations [35]. These collections, including the TIPSTER data [34], deserve revisiting with modern language modeling techniques. While the relevance criteria are different, these collections are still useful from a language understanding perspective. Particularly relevant are narrative-driven recommendation datasets for books [14] and points-of-interest [1], as the ground truth aligns with our definition of item utility. The difference, and thus limitation of these, lies in the nature of the narratives, which focus on ephemeral contextual information needs as opposed to standing preferences. Closest to our setting, Balog and Radlinski [4] solicited self-provided descriptions of liked and disliked items (movies) from users, along with specific item examples. This somewhat simplified version of our task also had recommendations made by humans from a small set of (60) candidate items. Nevertheless, the methodology can be adapted to facilitate dataset construction on a larger scale. Note that all the above are offline test collections, which focus on this task in isolation. We discuss desiderata for more holistic evaluation in Sec. 5.5.

5 CHALLENGES AND OPPORTUNITIES

In presenting scrutable recommendation, we touched upon a number of challenges. We return to some of these, suggesting a research agenda toward realizing the benefits of scrutable recommendation.

5.1 Generating Natural Language Summaries

5.1.1 Characteristics of NL Summaries. The first significant challenge is what type of terminology is optimal in the natural language summary? Past work has focused on attributes and social tags [5, 13], but this may not provide sufficient granularity or expressiveness of preferences. While text reviews, search queries, and crowdsourced critiques are closer to a narrative description, they usually do not express preferences with as much detail as an expert in a given domain might when summarizing a particular taste refinement. Transcripts of user-generated descriptions of preferences collected by Radlinski et al. [63] suggest a particular lack of formality and specificity in many user-generated descriptions. The same corpus illustrates that most people rarely referred to metadata frequently used to characterize movies in expert-written text, such as actor and director names. This raises the question of what *sort* of language would be best for summaries of user preferences? And, how can these summaries be made fluent, precise, and using the terminology that people would choose? Similarly, could meta-preferences—such as preferences for novelty or tolerance of repetition (say when listening to music)—be captured?

5.1.2 Cold-start Recommendations. New users provide a particularly impactful case for NL summaries. In cold start situations, users would, by definition, lack substantial input to provide to a recommendation algorithm. This begs the question of how to solicit useful information from them. The work by Bogers and Koolen [14] suggests that NL preference communication can require many requests, with clarification and refinement. How an algorithm would efficiently solicit preferences is unclear: Should a system even solicit, or would it be better for new users to simply write text? Past Wizard-of-Oz studies have first solicited examples, then explanations of these [63]. Recent work has also generated user summaries using additional information (e.g., age and occupation) along with a few reviews, or user information from auxillary domains [44, 45].

In contrast to the user cold-start setting, other challenges come up in characterizing new items: if a user indicates a preference for/against a particular item, a system would need to know how to generalize to other items. If a new item has few existing reviews or other details available, this task would be particularly challenging.

5.1.3 Cognitive Load. An overarching assumption has been that text is desirable, and mathematical representations are not. It assumes text is more scrutable and requires less cognitive effort to improve and control recommendations. A key question is whether it really has sufficiently low cognitive load. Indeed, while users may want to understand and control their user summary, approaches that require large effort are likely undesirable. Therefore, it is important to consider ways to reduce the cognitive load required to generate, understand, and update user summaries. One obvious way is to ensure that representations are relatively concise. Other methods may include consistent phrasing, or providing multiple phrasings so that users can pick those that appeal to them. Other attributes of text such as readability, complexity of language, linguistic forms, and choice of terminology may also influence cognitive load, and how to trade off these aspects with the informativeness of a summary may need to be addressed.

Indeed, just text may *not* be optimal: *multi-modal representations* may be desirable, such as text combined with other user-defined items that represent interests—e.g., images or even audio snippets in a music recommendation setting. Any representation that maintains scrutability but improves performance or reduces cognitive load is likely valuable.

5.2 Updating Natural Language Summaries

A special case of generating natural language summaries is algorithmically updating existing summaries given new evidence.

5.2.1 Managing Updates of Representation. A NL summary needs to adapt as user preferences change. This means the system must recognize changes in interests, and either automatically adjust the profile or request confirmation from the user. One possible avenue would be to maintain two parts of a profile: preferences directly expressed or confirmed by the user, as well as those inferred but not yet validated. Recommendations and explanations can then provide as a basis specific aspects of each, allowing for freshness, transparency, as well as control.

5.2.2 Recognizing New Contexts. A seemingly straightforward case would be if interactions are observed with items that do not match a profile at all. For instance, unexpected interactions with items tailored for young children. A question arises of when, and how, such preferences should be incorporated into a user model. Factoring this as a new context is likely beneficial, especially as many preferences are contextual (e.g., [72]). Such an effect was observed in the travel domain for point-of-interest recommendation in a new location [73]. However, while a new location might suggest a short-term context "when in Hawaii," a sequence of such observations over a longer time may suggest the user profile should actually capture a new context "when on vacation." Hence aggregating preferences that seemingly fall into different contexts is a further challenge.

5.2.3 Refining User Summaries. A second situation where a user profile is likely to evolve is when expertise changes. For instance, as someone explores music in a new genre, they may refine their taste to a sub-genre, a specific era, specific artists, and so forth. As such, they may choose to represent their preferences in generic language initially, refining it over time. Suggested refinements from a generative system are likely to be valuable here, and it is an open question as to how such refinements should be derived and represented so as to capture a person's interests most effectively.

5.2.4 Realizing Dislikes. A final aspect of updating NL summaries relates to the value of explicit *negative* preferences, as differentiated from a lack of preferences about items that a user may not know about. Negative user preferences discussed in [69] were modeled by Balog et al. [5], and it is natural to ask how these trade off with more details about positive preferences. The trade-off with scrutability is particularly pertinent if the role of the recommender is to present broad recommendations, with users valuing serendipity and diversity [59, 90].

5.3 Utilizing Natural Language Summaries

We move on to taking advantage of NL summaries for recommendation, expanding on Sec. 4 to focus on more open-ended challenges.

5.3.1 Interpreting Language Robustly. When generating recommendations from NL summaries it is important that a system reliably interprets nuanced concepts. While large language models have made tremendous strides in language understanding, ensuring they reliably map from terms to concepts is critical. Approaches like universal sentence encoders [20] provide one way to map from text to vector input suitable for traditional recommender algorithms, but text understanding must also be robust to slight variations in wording, verifying that small changes in wording do not affect recommendations significantly, as this would limit the usefulness of users scrutinizing them. Noting that descriptions may be ambiguous (e.g., "Star Wars," may be a movie, or a series of movies, or a fictional universe with different types of content), errors in understanding preferences when there is little text for any given individual have a large effect. A specific challenge is that of subjectivity, where the same word means different things to different people. When such language is used, even a user-inspected description may be misinterpreted [64]. Recognizing *when* there is potential for misunderstanding, *resolving* it and even *expressing assumptions* about interpretations explicitly are ways this could be addressed.

5.3.2 Explaining Recommendations. Recently Lyu et al. [50] showed that human recommenders spend the most significant part of their time (38%) explaining recommendations, suggesting these serve a critical role in the recommendation process. While past work showed that useful explanations can be generated from non-language-based models [89], NL models that have been *validated by the user* offer an even stronger opportunity: it may be possible to explain recommendations using the *user's* preferred terminology and phrasing. For instance, describing a movie as *funny* may be desirable, yet different people may mean different things by this term [6]. Using the user's specific meaning in explanations may be beneficial. Similarly, it may be possible to personalize to the user's particular language preferences: some people may prefer more or less formal explanations, longer or shorter explanations, explanations that bring attention to different aspects of items, and so forth.

Explanations can also offer recourse if a match does not exist from the user's perspective but exists from the system's perspective. For instance, an explanation may bring attention to an element of someone's profile that is not specific enough to exclude non-relevant results. As yet, approaches to enable such interaction with explanations have received limited attention (e.g., [65]).

5.3.3 Practical Utility. Commonly, recommendation approaches are evaluated based on accuracy [67]. However, as noted above, *novel* items often provide most value to users [22]. A natural challenge with short text summaries of user's interests is that they could not record every item the person already knows or has previously received as a recommendation: every movie watched, every song listened to, etc. To satisfy a desire for novel content, it may be valuable for a system to have some side information. The precise form this should take, and how to process this information while respecting the goals of transparency and user control is an important challenge.

In particular, it would be important to evaluate if observational information used in a limited manner (e.g., to filter out repeated items) is preferred over it being used for recommendation. An open challenge is the development of metrics that then allow the utility of such additional side information to be traded off against additional complexities in how users may scrutinize and modify their information. It is also likely there are different types of additional information that could be consumed, and that could impact results in different ways, so as to combine the strengths of different recommendation approaches most effectively.

5.4 Optimization

For generating natural language summaries, and using them for recommendation, optimization is an important challenge.

5.4.1 Trading off Competing Goals. As mentioned earlier, there are many goals for the natural language descriptions (scrutable, concise, fluent, useful for recommendation) as well as for actual recommendations (accurate, novel, diverse). These competing goals need to draw on techniques that jointly optimize many metrics. In particular, the trade-off between the goals for the user summary and those for recommendations are unclear. We hypothesize that different people may have different priorities for recommendation quality versus scrutability, suggesting that a single learned model should permit adjusting this on a per-person basis.

5.4.2 End-to-end Optimization. Given a suitable combined loss, the next question is how to optimize it. While we have described a two part process for recommendation, recent work on β variational autoencoders has shown that partitioning may not be necessary: it may be possible to train a recommendation algorithm for both high accuracy and desirable properties of the user encoding (such as controllability) in an end-to-end manner. E.g., Nema et al. [54] showed how to train a β-VAE so that particular attributes can be explicitly controlled in recommendations. It is possible that a recommendation model could similarly learn a NL user profile end-to-end, encoding desirable properties in the optimization loss.

5.5 Evaluation

As a final set of challenges, we turn to *end-to-end* evaluation (component-level evaluation being discussed in Sec. 3.2 and 4.2).

5.5.1 Task-Specific Time/Data Efficiency. When a user engages with a NL summary—be it creating or modifying one—it would be important to understand the concrete utility to the user. For example, a NL summary of a given length may take a given amount of time and effort to produce. How does this compare to the number of traditional ratings that a person could produce in the same amount of time? Is there a relationship that a given number of language tokens is equivalent to a specific number ratings, in terms of recommendation performance?

Notably, when there is a shift in user preferences, the effect of adding or removing a sentence from the NL summary could be contrasted against the number of ratings the user would need to change for the same effect. Perhaps recommendations from a natural language model may be poorer than state-of-the-art rating-based recommender systems in a non-cold start scenario, yet the ease of curation may close this gap for users who choose to improve their model in language rather than by changing ratings.

5.5.2 Coverage versus Specificity. Suppose a description could either perfectly describe a fraction of the user's interests, or cover all their preferences but less well. An evaluation must ask: Which is better? Top-N recommendations typically prefer the former, but it is more difficult to quantify the latter: What fraction of all the user's interests are represented in the output of a recommendation algorithm?

5.5.3 Types of Utility and Cognitive Effort. Overall, it is important to observe that there are two types of utility: recommendation quality and language utility. From a system perspective, specific types of description elements may have the strongest effect on recommendations (say, examples of attributes that do or do not appeal to the user). From a user perspective, this may be different—it may be unnatural or difficult for someone to provide text with the right sort of language. On the other hand, other types of natural language input may be less useful to the system but easier for a person to generate or edit (say, free-form descriptions of items that appeal and do not appeal). As such, evaluation methodologies that incorporate cognitive effort as well as time to improve are important to develop. As discussed in Sec. 5.2, there may be approaches that assist in this curation process such as suggesting phrases, flagging ambiguity, or asking explicit questions.

Similarly, as scrutable NL-based recommendations permit users to edit their own preference summaries directly, these should have a direct and immediate impact on the recommendations that they receive. Possible edits could be simulated to measure how different types of edits affect recommendations.

Table 2: Few-shot prompt template used to turn a large language model into a recommender.

```
"contemplative, trippy, mind-bending or artsy sci-fi movies":
"2001: A Space Odyssey" - a classic sci-fi movie by Stanley Kubrick, with lots of trippy and artsy moments.
"Primer" - a mind-bending time travel movie where timelines constantly intersect and overlap.
"Arrival" - a contemplative sci-fi movie about first contact.

"cool sci-fi settings that are somewhat dystopic":
"Blade Runner" - a film noir sci-fi, where rogue AIs are hunted down.
"District 9" - aliens visit Earth, and are treated as second-class citizens.
"Gattaca" - a near-future dystopia about genetics determining your fate.

"funny and non-traditional superhero movies":
"The Lego Movie" - basically a superhero movie, full of humor and doesn't take itself too seriously.
"Thor: Ragnarok" - the silliest of all the Thor movies, with some great secondary characters.
"Deadpool" - the humor is over the top and obscene, with a relentlessly funny narrator.

"<NL-SUMMARY>":
```

6 FEASIBILITY STUDY

We now present a small case study highlighting how a large language model can be used for movie recommendation based on NL summaries. In particular, we use LaMDA, a decoder-only transformer-based, 137B parameter language model [76]. It is trained only to predict the next token in a large generic text corpus. The results from this model are similar to other large language models such as GPT-3 [17].

The model is adapted for recommendation by configuring it using "few-shot" prompt templates [17]: Table 2 presents the exact input given to it, where the placeholder `<NL-SUMMARY>` was replaced by the user's NL summary. The model is then simply used to predict the next token, repeatedly within a top-K decoding strategy. The

Table 3: Example NL summary inputs with the corresponding generated output text that includes recommendations and explanations that reference the NL summary.

NL Rep.: movies where the murderer is involved from the beginning **"The Usual Suspects"** - a movie where the narrator is unreliable, and the murderer is involved from the beginning. **"Se7en"** - a detective story about a serial killer, where the killer is revealed from the beginning. **"Scream"** - a classic horror movie where the killer is revealed early on. **"Memento"** - a movie about a man with short-term memory loss, who is trying to find his wife's killer.
NL Rep.: space exploration movies that are funny **"Guardians of the Galaxy"** - a team of intergalactic misfits, who are funny and have a lot of heart. **"The Martian"** - a great adaptation of the book by Andy Weir, with a great cast. **"Spaceballs"** - a parody of Star Wars and other sci-fi movies. **"The Hitchhiker's Guide to the Galaxy"** - a funny sci-fi movie that parodies space exploration.
NL Rep.: animated movies with creative and interesting fight scenes **"Akira"** - a classic anime movie about a boy with telekinetic powers. **"The Incredibles"** - a family of superheroes with some of the best fight scenes you'll see. **"Kung Fu Panda"** - a kung fu movie with a panda as the protagonist. **"Sword of the Stranger"** - a samurai movie, with creative fight scenes and a bittersweet ending.

output from a single run for three example NL summaries is shown in Table 3, showing the first four recommendations for each.[4]

Given that no model fine-tuning has been applied, and the model was primed with only three examples of NL summaries paired with three recommendations and brief explanations, it is impressive that it gives both appropriate recommendations and personalized explanations. For example, referencing "creative fight scenes" in animated movies, instead of their animation style when the user expressed interest in "creative and interesting fight scenes." When contrasted to entering the NL summary as queries into popular search engines, we observe search engines provide links to pages of lists, e.g., "space exploration movies that are funny," capturing aspects of the NL query but much less specifically than the language model results, and lacking explanations. In some cases, e.g., "space exploration movies that are funny," search engines do interpret the query as genre constraints for a specialised movie search, in which case there is a reasonable overlap with the NL recommender's suggestions, but again without explanation.

While Table 3 may give the initial appearance of having solved the NL recommendation problem, there are significant limitations. Among those described in Sec. 5, we particularly observe: (1) The NL summaries above are short, focusing on a single aspect, rather than holistic summaries of real people's broad interests. System performance with more realistic user descriptions matters more (cf. Sec. 5.1.1). (2) If users prefer novel recommendations, the recommendations here may not respect this (cf. Sec. 5.3.3). (3) Language models can contain hard to identify unintended biases. For example, we observe that the movies recommended by this model may skew towards popular movies in the United States, and are less accurate if queries include specific constraints like precise release decades (cf. Sec. 5.3.1). (4) Language models have been observed to hallucinate. For example, the first example in Table 3 highlights "Se7en" as a movie with a killer revealed in the beginning, but the killer is revealed much later in the film (cf. Sec. 5.3.2).

[4]We note that while the few-shot templates were crafted, the NL summaries and outputs were not cherry-picked.

7 CONCLUSIONS

We argue that recent advances in natural language generation and understanding make it possible to envision a fundamentally new approach to personalized search and recommendation, one where systems' knowledge of users is encoded in natural language. In providing a process by which people can review and update their representations, it could significant improve transparency, user control, and, ultimately, quality of recommendations. This paper presents a breakdown of the end-to-end task of scrutable recommendation into more tractable steps and identifies both promising recent advances as well as open challenges for the broader research community to address.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers as well as Alex Beutel for their useful feedback on this paper.

REFERENCES

[1] Jafar Afzali, Aleksander Mark Drzewiecki, and Krisztian Balog. 2021. POINTREC: A Test Collection for Narrative-Driven Point of Interest Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. 2478–2484.

[2] Azhar Alhindi, Udo Kruschwitz, Chris Fox, and M-Dyaa Albakour. 2015. Profile-Based Summarisation for Web Site Navigation. *ACM Trans. Inf. Syst.* 33, 1, Article 4 (feb 2015), 39 pages.

[3] Jaime Arguello, Bogeum Choi, and Robert Capra. 2018. Factors Influencing Users' Information Requests: Medium, Target, and Extra-Topical Dimension. *ACM Trans. Inf. Syst.* 36, 4 (July 2018), 41:1–41:37.

[4] Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 329–338.

[5] Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. 2019. Transparent, Scrutable and Explainable User Models for Personalized Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. 265–274.

[6] Krisztian Balog, Filip Radlinski, and Alexandros Karatzoglou. 2021. On Interpretation and Measurement of Soft Attributes for Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. 890–899.

[7] Nicholas J. Belkin and W. Bruce Croft. 1992. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Commun. ACM* 35, 12 (dec 1992), 29–38.

[8] Moshe Ben-Akiva, Takayuki Morikawa, and Fumiaki Shiroishi. 1992. Analysis of the Reliability of Preference Ranking Data. *J. Bus. Res.* 24, 2 (1992), 149–164.

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. 610–623.

[10] Michael Bendersky and W. Bruce Croft. 2008. Discovering Key Concepts in Verbose Queries. In *Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. 491–498.

[11] Michael Bendersky and W. Bruce Croft. 2009. Analysis of Long Queries in a Large Scale Search Log. In *Proceedings of the 2009 Workshop on Web Search Click Data (WSCD '09)*. 8–14.

[12] Michael Bendersky, Donald Metzler, and W. Bruce Croft. 2010. Learning Concept Importance Using a Weighted Dependence Model. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM '10)*. 31–40.

[13] Toine Bogers. 2018. *Tag-Based Recommendation.* Springer International Publishing, Cham, 441–479.

[14] Toine Bogers and Marijn Koolen. 2017. Defining and Supporting Narrative-Driven Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems (RecSys '17)*. 238–242.

[15] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. 2012. TasteWeights: A Visual Interactive Hybrid Recommender System. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. 35–42.

[16] Ronen I. Brafman, Carmel Domshlak, Solomon Eyal Shimony, and Y. Silver. 2006. Preferences over Sets. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference*. 1101–1106.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901.

[18] Pablo Castells, Neil J. Hurley, and Saul Vargas. 2015. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook* (2nd ed.), Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). 881–918.

[19] Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2021. Evaluation of Text Generation: A Survey. arXiv:cs.CL/2006.14799

[20] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP '18)*. 169–174.

[21] Shuo Chang, F. Maxwell Harper, and Loren Terveen. 2015. Using Groups of Items for Preference Elicitation in Recommender Systems. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. 1258–1269.

[22] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. In *The World Wide Web Conference (WWW '19)*. 240–250.

[23] Marie desJardins, Eric Eaton, and Kiri L. Wagstaff. 2006. Learning User Preferences for Sets of Objects. In *Proceedings of the 23rd International Conference on Machine Learning (ICML '06)*. 273–280.

[24] Simon Dooms, Toon De Pessemier, and Luc Martens. 2014. Improving IMDb Movie Recommendations with Interactive Settings and Filters. In *Poster Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)*.

[25] Frederico Araújo Durão and Peter Dolog. 2009. A Personalized Tag-Based Recommendation in Social Web Systems. In *Proceedings of the Workshop on Adaptation and Personalization for Web 2.0*.

[26] Michael D Ekstrand and Martijn C Willemsen. 2016. Behaviorism is Not Enough: Better Recommendations through Listening to Users. In *Proceedings of the 10th ACM conference on recommender systems (RecSys '16)*. 221–224.

[27] Claudiu S Firan, Wolfgang Nejdl, and Raluca Paiu. 2007. The Benefit of Using Tag-Based Profiles. In *2007 Latin American Web Conference (LA-WEB '07)*. 32–41.

[28] Jonathan Furner. 2002. On Recommending. *J. Assoc. Inf. Sci. Technol.* 53 (2002), 747–763.

[29] Jean Garcia-Gathright, Brian St. Thomas, Christine Hosey, Zahra Nazari, and Fernando Diaz. 2018. Understanding and Evaluating User Satisfaction with Music Discovery. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. 55–64.

[30] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. *Int. J. Hum.-Comput. Stud.* 72, 4 (April 2014), 367–382.

[31] Manish Gupta and Michael Bendersky. 2015. Information Retrieval with Verbose Queries. *Found. Trends Inf. Ret.* 9, 3–4 (2015), 209–354.

[32] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris van Hoboken. 2019. Designing for the Better by Taking Users into Account: A Qualitative Evaluation of User Control Mechanisms in (News) Recommender Systems. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*. 69–77.

[33] Negar Hariri, Bamshad Mobasher, and Robin Burke. 2015. Adapting to User Preference Changes in Interactive Recommendation. In *Proceedings of the 24th International Conference on Artificial Intelligence (IJCAI '15)*. 4268–4274.

[34] Donna Harman. 1992. The DARPA Tipster Project. In *ACM SIGIR Forum*, Vol. 26. 26–28.

[35] Donna Harman. 1993. Document Detection Data Preparation. In *TIPSTER TEXT PROGRAM: PHASE I: Proceedings of a Workshop held at Fredricksburg, Virginia, September 19-23, 1993*. 17–31.

[36] F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. 2015. Putting Users in Control of Their Recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems (RecSys '15)*. 3–10.

[37] C. B. Hensley. 1963. Selective Dissemination of Information (SDI): State of the Art in May, 1963. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference (AFIPS '63 (Spring))*. 257–262.

[38] Yoshinori Hijikata, Yuki Kai, and Shogo Nishida. 2012. The Relation between User Intervention and User Satisfaction for Information Recommendation. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing (SAC '12)*. 2002–2007.

[39] Dietmar Jannach, Sidra Naveed, and Michael Jugovac. 2017. User Control in Recommender Systems: Overview and Interaction Challenges. In *E-Commerce and Web Technologies - 17th International Conference, Revised Selected Papers (EC-Web '16)*. 21–33.

[40] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of Personal Characteristics on Music Recommender Systems with Different Levels of Controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. 13–21.

[41] Jinyoung Kim, Hyungjin Kim, and Jung-hee Ryu. 2009. TripTip: A Trip Planning Service with Tag-based Recommendation. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems (CHI EA '09)*. 3467–3472.

[42] Bart P. Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. Inspectability and Control in Social Recommenders. In *Proceedings of the Sixth ACM Conference on Recommender Systems (RecSys '12)*. 43–50.

[43] Lorenz Kuhn and Carsten Eickhoff. 2016. Implicit Negative Feedback in Clinical Information Retrieval. In *Proceedings of the SIGIR 2016 Medical Information Retrieval Workshop (MedIR)*.

[44] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. 2008. Addressing Cold-Start Problem in Recommendation Systems. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication (ICUIMC '08)*. 208–211.

[45] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. MeLU: Meta-Learned User Preference Estimator for Cold-Start Recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19)*. 1073–1082.

[46] Liu Leqi, Fatma Kilinc-Karzan, Zachary C. Lipton, and Alan L. Montgomery. 2021. Rebounding Bandits for Modeling Satiation Effects. In *Advances in Neural Information Processing Systems 34 pre-proceedings (NeurIPS '21)*.

[47] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP '21)*. 3045–3059.

[48] Hans Peter Luhn. 1958. A Business Intelligence System. *IBM J. Res. Dev.* 2, 4 (Oct. 1958), 314–319.

[49] Kai Lukoff, Ulrik Lyngs, Himanshu Zade, J. Vera Liao, James Choi, Kaiyue Fan, Sean A. Munson, and Alexis Hiniker. 2021. How the Design of YouTube Influences User Sense of Agency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Article 368.

[50] Shengnan Lyu, Arpit Rana, Scott Sanner, and Mohamed Reda Bouadjenek. 2021. A Workflow Analysis of Context-driven Conversational Recommendation. In *Proceedings of the Web Conference 2021 (WWW '21)*. 866–877.

[51] 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)*. 25–34.

[52] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Making Recommendations Better: An Analytic Model for Human-Recommender Interaction. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. 1103–1108.

[53] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. 2020. WT5?! Training Text-to-Text Models to Explain their Predictions. arXiv:cs.CL/2004.14546

[54] Preksha Nema, Alexandros Karatzoglou, and Filip Radlinski. 2021. Disentangling Preference Representations for Recommendation Critiquing with β–VAE. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM '21)*. 1356–1365.

[55] Thao Ngo, Johannes Kunkel, and Jürgen Ziegler. 2020. Exploring Mental Models for Transparent and Controllable Recommender Systems: A Qualitative Study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*. 183–191.

[56] John O'Donovan and Barry Smyth. 2005. Trust in Recommender Systems. In *Proceedings of the 10th International Conference on Intelligent User Interfaces (IUI '05)*. 167–174.

[57] Jiaul H. Paik and Douglas W. Oard. 2014. A Fixed-Point Method for Weighting Terms in Verbose Informational Queries. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. 131–140.

[58] Javier Parapar and Filip Radlinski. 2021. Diverse User Preference Elicitation with Multi-Armed Bandits. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM '21)*. 130–138.

[59] Javier Parapar and Filip Radlinski. 2021. Towards Unified Metrics for Accuracy and Diversity for Recommender Systems. In *Proceedings of the Fifteenth ACM Conference on Recommender Systems (RecSys '21)*. 75–84.

[60] Denis Parra and Peter Brusilovsky. 2015. User-controllable Personalization: A Case Study with SetFusion. *Int. J. Hum. Comput.* 78 (2015), 43–67.

[61] Javier Parra-Arnau, Jagdish Prasad Achara, and Claude Castelluccia. 2017. MyAdChoices: Bringing Transparency and Control to Online Advertising. *ACM Trans. Web* 11, 1 (2017), 1–47.

[62] Fabíola S. F. Pereira, João Gama, Sandra de Amo, and Gina M. B. Oliveira. 2018. On Analyzing User Preference Dynamics with Temporal Social Networks. *Mach. Learn.* 107 (2018), 1745–1773.

[63] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL '19)*. 353–360.

[64] Filip Radlinski, Craig Boutilier, Deepak Ramachandran, and Ivan Vendrov. 2022. Subjective Attributes in Conversational Recommendation Systems: Challenges and Opportunities. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI '22)*.

[65] Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David Lagnado, and Francesca Toni. 2021. Argumentative Explanations for Interactive Recommendations. *Artif. Intell.* 296 (2021), 103506.

[66] Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating Pretrained Language Models for Graph-to-Text Generation. arXiv:cs.CL/2007.08426

[67] Francesco Ricci, Lior Rokach, and Bracha Shapira. 2015. *Recommender Systems Handbook* (2nd ed.). Springer.

[68] Noveen Sachdeva and Julian McAuley. 2020. How Useful Are Reviews for Recommendation? A Critical Review and Potential Improvements. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. 1845–1848.

[69] Shilad Sen, F. Maxwell Harper, Adam LaPitz, and John Riedl. 2007. The Quest for Quality Tags. In *Proceedings of the 2007 International ACM Conference on Supporting Group Work (GROUP '07)*. 361–370.

[70] Shilad Sen, Shyong K. Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F. Maxwell Harper, and John Riedl. 2006. Tagging, Communities, Vocabulary, Evolution. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work (CSCW '06)*. 181–190.

[71] Rashmi Sinha and Kirsten Swearingen. 2002. The Role of Transparency in Recommender Systems. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems (CHI EA '02)*. 830–831.

[72] Kostas Stefanidis, Evaggelia Pitoura, and Panos Vassiliadis. 2011. Managing Contextual Preferences. *Inf. Syst.* 36, 8 (2011), 1158–1180.

[73] Ke Sun, Tieyun Qian, Tong Chen, Yile Liang, Quoc Viet Hung Nguyen, and Hongzhi Yin. 2020. Where to Go Next: Modeling Long- and Short-Term User Preferences for Point-of-Interest Recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 01 (2020), 214–221.

[74] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '18)*. 235–244.

[75] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. 2005. Personalizing Search via Automated Analysis of Interests and Activities. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*. 235–244.

[76] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. arXiv:cs.CL/2201.08239

[77] Nava Tintarev and Judith Masthoff. 2012. Evaluating the Effectiveness of Explanations for Recommender Systems. *User Model. User-adapt. Interact.* 22 (oct 2012), 399–439.

[78] Nava Tintarev and Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook* (2nd ed.), Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor (Eds.). Springer, 353–382.

[79] Jan Trienes and Krisztian Balog. 2019. Identifying Unclear Questions in Community Question Answering Websites. In *Proceedings of the 41th European Conference on Advances in Information Retrieval (ECIR '19)*. 276–289.

[80] Daniel Valcarce, Alejandro Bellogín, Javier Parapar, and Pablo Castells. 2018. On the Robustness and Discriminative Power of Information Retrieval Metrics for Top-N Recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. 260–268.

[81] Jesse Vig, Shilad Sen, and John Riedl. 2009. Tagsplanations: Explaining Recommendations Using Tags. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)*. 47–56.

[82] Ellen M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*.

[83] Ellen M. Voorhees and Donna K. Harman (Eds.). 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.

[84] Claudia Wagner, Philipp Singer, Markus Strohmaier, and Bernardo A. Huberman. 2014. Semantic Stability in Social Tagging Streams. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. 735–746.

[85] Bishan Yang, Nish Parikh, Gyanit Singh, and Neel Sundaresan. 2014. A Study of Query Term Deletion Using Large-Scale E-commerce Search Logs. In *Proceedings of the 36th European Conference on Information Retrieval Research: Advances in Information Retrieval (ECIR '14)*. 235–246.

[86] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational Information Seeking. arXiv:cs.IR/2201.08808

[87] Yury Zemlyanskiy, Sudeep Gandhe, Ruining He, Bhargav Kanagal, Anirudh Ravula, Juro Gottweis, Fei Sha, and Ilya Eckstein. 2021. DOCENT: Learning Self-Supervised Entity Representations from Large Document Collections. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (EACL '21)*. 2540–2549.

[88] Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (apr 2004), 179–214.

[89] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Ret.* 14, 1 (2020), 1–101.

[90] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining (WSDM '12)*. 13–22.

[91] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (NAACL '18)*. 15–20.

[92] Le Zhao and Jamie Callan. 2010. Term Necessity Prediction. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM '10)*. 259–268.