Factors Influencing Users' Information Requests: Medium, Target, and Extra-Topical Dimension

JAIME ARGUELLO, BOGEUM CHOI, and ROBERT CAPRA, School of Information and Library Science, University of North Carolina at Chapel Hill

We report on a crowdsourced study that investigated how two factors influence the way people formulate information requests. Our first factor, *medium*, considers whether the request is produced using text or voice. Our second factor, *target*, considers whether the request is intended for a search engine or a human intermediary (i.e., someone who will search on the user's behalf). In particular, we study how these two factors influence the way people formulate requests in situations where the information need has a specific type of *extra-topical* dimension (i.e., a type of constraint that is independent from the information need's topic). We focus on six extra-topical dimensions: (1) domain knowledge, (2) viewpoint, (3) experiential, (4) venue location, (5) source location, and (6) temporal. The extra-topical dimension was manipulated by giving participants carefully constructed search tasks. We analyzed a large number of information requests produced by study participants, and address three research questions. We study the effects of our two factors (medium and target) on (RQ1) participants' perceptions about their own information requests, (RQ2) the different characteristics of their information requests (e.g., natural language structure, retrieval performance), and (RQ3) participants' strategies for requesting information when the search task has a specific type of extra-topical dimension. Our results found that both factors influenced participants' perceptions about their own information requests, the characteristics of participants' requests, and the strategies adopted by participants to request information matching the extra-topical dimension. Our results have implications for future research on methods that can harness (rather than ignore) extra-topical query terms to retrieve relevant information.

CCS Concepts: • **Information systems** → **Search interfaces**;

Additional Key Words and Phrases: Information requests, query formulation, spoken search, relevance criteria, search performance

ACM Reference format:

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, Article 41 (July 2018), 37 pages. <https://doi.org/10.1145/3209624>

1 INTRODUCTION

In recent years, “conversational” search interfaces have become increasingly common. Such systems include mobile and desktop personal assistants such as Siri and Cortana, home assistants such as Amazon's Alexa and Google Home, and automobile assistants such as Nuance's Dragon Drive,

This work was supported by NSF grant IIS-1451668. Any opinions, findings, conclusions, and recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the sponsor.

Authors' addresses: J. Arguello, B. Choi, and R. Capra, School of Information and Library Science, University of North Carolina at Chapel Hill, 216 Lenoir Drive, CB #3360, 100 Manning Hall, Chapel Hill, NC 27599-3360; emails: jarguello@unc.edu, choiboge@live.unc.edu, rcapra@unc.edu.

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

© 2018 ACM 1046-8188/2018/07-ART41 \$15.00

<https://doi.org/10.1145/3209624>

which provides point-of-interest search capabilities. These novel search interfaces are different from traditional search interfaces in two respects. First, they are able to accept voice (vs. textual) information requests from users. Second, they aim to engage users through more human-like interactions. For example, mobile and desktop assistants are currently able to perform anaphora resolution based on previous searches (e.g., “Who is president of France?”, “How old is he?”).

As search systems continue to evolve and become more conversational and naturalistic, we need to understand how their differences from traditional interfaces may influence the way users request information. Prior work has investigated the differences between voice and textual queries [8–10, 12, 26, 34]. However, with respect to conversational search interfaces, open questions remain. For example, as systems support more naturalistic interactions, will users increase their expectations about the system’s ability to understand and respond to requests? Will users change the way they formulate requests? Will users contribute *more* information about their information need or task at hand? Will this additional information actually produce better search results? If not, why not?

In this article, we investigate how two different factors influence the way people request information. Our first factor, referred to as the *medium* of the information request, considers whether the request is produced using text or voice. Our second factor, referred to as the *target* of the information request, considers whether the request is intended for a search engine or a human search intermediary (i.e., someone who will search on the user’s behalf). Our goal in the “human intermediary” condition was to study information requests in the *extreme* case where users have a very high expectation about the system’s ability to understand and respond to requests.

In addition to investigating our two factors of medium and target, we focus on information needs that have a specific type of *extra-topical* dimension. It has long been recognized that relevance from a user’s perspective is a multidimensional construct, involving different factors other than the *topic* of the information need [24]. In this article, we focus on *six* extra-topical dimensions: (1) the user wants information for a domain novice or expert (*domain knowledge*); (2) the user wants information supporting a particular point of view (*viewpoint*); (3) the user wants information from someone with first-hand experience (*experiential*); (4) the user wants venue information from a specific location (*venue location*); (5) the user wants information for a specific geographical perspective (*source location*); and (6) the user wants information from a specific time period (*temporal*). A large body of prior work has investigated the different types of extra-topical criteria that influence users’ relevance decisions [4–7, 21, 22, 25, 27–29, 31–33]. However, less research has aimed at understanding how different extra-topical dimensions of the information need may influence the way people request information from a system [15–17]. Our research fits this second category.

In this article, we analyze a large number of information requests gathered using a crowdsourced study, and we address the following three research questions:

RQ1: In our first research question, we investigate how our two factors of medium and target influence users’ perceptions about their own information request. Specifically, we focus on users’ perceived level of difficulty in producing the information request, their level of satisfaction with the request, and their level of confidence that the request will yield high-quality search results.

RQ2: In our second research question, we investigate how our two factors of medium and target influence the characteristics of users’ information requests. We address this question from two perspectives. First, following prior research on the differences between voice and textual queries [12], we consider different surface-level characteristics such as the length of the information request (in words), part-of-speech composition, and level of grammatical complexity. Second, we compare the retrieval effectiveness of information requests produced under different conditions.

RQ3: In our third and final research question, we investigate how our two factors of medium and target influence the way people request information when they have a specific type of

extra-topical preference. We focus on the six extra-topical dimensions mentioned above. We address this question from three perspectives. First, we investigate the different strategies used by participants to request information when they have a specific type of extra-topical preference. Second, we consider the retrieval performance of different strategies. Finally, we investigate whether certain strategies are more common depending on the medium and/or target condition. For example, do people use more elaborate or complex strategies when producing voice requests versus text requests? Do people use more complex strategies when requesting information from a human versus a search engine? Ultimately, we are interested in whether our two factors of medium and target influence people to adopt strategies that impact *retrieval performance* given a current state-of-the-art retrieval system.

To investigate our three research questions, we conducted a large-scale crowdsourced study. Participants were given carefully constructed search tasks and asked to produce requests. RQ1 was investigated using posttask questionnaire data. RQ2 was investigated using different text analysis methods and by issuing our participants' requests to a commercial web search API. RQ3 was investigated through extensive qualitative analysis of participants' requests.

2 RELATED WORK

Our research builds on three areas of prior work: (1) understanding the differences between voice and textual queries, (2) understanding the different extra-topical relevance criteria that are important to users, and (3) understanding how different extra-topical criteria may influence the way people request information from a system.

2.1 Understanding Voice Queries

A large body of prior work has focused on understanding the differences between voice and textual queries, in terms of both their characteristics and retrieval performance. Furthermore, prior work has also investigated algorithms for automatically *altering* voice queries to improve retrieval performance.

Voice versus Textual Queries: Early work by Du and Crestani [8–10] explored the differences between voice and textual queries for the *same* information needs (based on TREC topics). With and without stopwords, voice queries were found to be longer than textual queries in both English [9] and Mandarin [10]. In English, voice queries also had more varied language—the total number of unique terms from all participants for the *same* information need was greater. In a follow-up analysis of the same English data, Crestani and Du [8] found three trends: (1) voice queries had fewer nouns and more of other parts of speech; (2) voice queries had *lower* retrieval effectiveness; and (3) both query types performed *equally* well after removing all terms except for nouns and adjectives. Additionally, a qualitative analysis found that voice queries were more likely to contain *extra-topical* language describing the desired document genre (e.g., news items, editorial comments, feature articles). The authors hypothesized that such terms decrease retrieval performance because they do not appear in relevant documents.

Prior work has also investigated the differences between voice and textual queries in the mobile search domain [12, 26, 34]. Schalkwyk et al. [26] compared voice and textual mobile queries and found three main trends: (1) mobile voice queries were associated with less interaction with the search results; (2) voice queries were more likely to be about “on the go” topics and less likely to be about sensitive topics such as adult themes and health; and (3) voice queries were found to be *shorter* than textual queries (possibly due to the “on the go” topics associated with voice queries). Yi and Maghoul [34] analyzed 79K spoken mobile queries from 2010 and found voice queries to be slightly longer than textual queries and, consistent with the previous paper [26], more often associated with topics requiring less interaction (i.e., good abandonment). More recently, Guy [12]

compared 500K spoken versus 500K textual queries from 2015. Voice queries were found to be longer than textual queries and had grammatical structures closer to natural language—voice queries had more *wh*-words, more parts of speech other than nouns, and deeper parse trees (indicating more full-sentence structures). Furthermore, consistent with previous work, spoken queries were more often associated with “quick search” topics requiring less interaction (e.g., voice queries more often triggered entity cards and returned multimedia content). Finally, spoken queries had more words that are easy to pronounce and difficult to spell, while textual queries had more words that are easy to type and difficult to speak (e.g., abbreviations and calendar years).

In this article, we replicate some of the findings from prior work (e.g., voice queries are more grammatically complex than textual queries [12]). Additionally, we investigate people’s perceptions about their own requests in different medium conditions (text vs. voice). Finally, we also consider the differences between requests intended for a search system versus a human intermediary. For example, we consider whether voice requests intended for a human have more natural language structure than voice requests intended for a search system.

Processing Verbose Voice Queries: As previously mentioned, prior work has found that while *unmodified* voice queries are longer than textual queries, they also yield worse retrieval performance [8]. A large body of prior work has focused on automatically reducing verbose textual queries in order to improve performance, for example, by deleting [18] or reweighting query terms [20] (see Gupta and Bendersky [11] for a review). Arguello et al. [2] focused on automatically reducing *voice* queries in order to improve performance. The authors proposed a machine-learning approach for predicting a subquery’s retrieval performance using query difficulty predictors as input features. The authors also analyzed the types of terms that are typically dropped from better-performing subqueries. Based on this analysis, problematic query terms included speech recognition errors; terms not central to the information need (e.g., “find,” “information”); and terms associated with extra-topical dimensions of the information need, such as temporal preferences (e.g., “current,” “recent,” “history”) and viewpoint preferences (e.g., “cons,” “negative,” “pros”). The last category suggests that search algorithms that rely on term matching are not always able to effectively handle extra-topical language in a query.

We extend this prior work by *explicitly* manipulating search tasks based on different extra-topical dimensions. Furthermore, through a qualitative analysis of participants’ requests, we investigate the strategies adopted by participants while requesting information matching a particular extra-topical constraint. Finally, we study whether people adopt strategies that yield better or worse retrieval performance based on the medium and target condition.

2.2 Relevance Criteria

In this article, we investigate how people request information in the presence of six different *types* of extra-topical preferences: (1) domain knowledge, (2) viewpoint, (3) experiential, (4) venue location, (5) source location, and (6) temporal preferences. An important question is: How do these six extra-topical dimensions relate to the different criteria that are important to people when searching for information? Are our six extra-topical dimensions grounded on criteria that are important to searchers?

It has long been recognized that relevance is a multidimensional concept. In his 2007 paper, Saracevic [24] summarized his and others’ research on relevance and argued that relevance from a user’s perspective is influenced by different *extra-topical* factors, including cognitive factors (e.g., what the user knows), affective factors (e.g., what the user feels or believes), and situational factors (e.g., the appropriateness of the information in helping a user complete the task at hand).

A large body of prior work has studied the different criteria people use when assessing the relevance of information for a specific need [4–7, 21, 22, 25, 27–29, 31–33]. Much of this research has

used qualitative methods to analyze people's reasons or justifications for deciding that a specific information item (e.g., a document, bibliographic record, or resource) is relevant or not relevant to their need. People's reasons or justifications have been gathered using a wide range of methods, including diary entries [6], document annotations [21], posttask interview responses [4, 7, 27], concurrent think-aloud comments [25, 29, 31, 32], and questionnaire responses produced by participants when deciding to "save" a document [33]. As one example, Maglaughlin and Sonnenwald [21] conducted a study with 12 social science students who were asked to produce relevance judgments and annotations for scientific articles given to them after a reference interview. Participants were then interviewed about their relevance assessments and annotations, and a qualitative coding of this interview data revealed 29 different relevance criteria cited by participants, including the article's credibility, depth/scope, novelty, understandability, temporal focus, and affective appeal.

Next, we discuss how our six extra-topical dimensions relate to relevance criteria observed in previous studies. Several studies observed domain knowledge as being one of the relevance criteria mentioned by participants (under labels such as "background knowledge" or "familiarity") [4, 22, 25, 28, 33]. Moreover, several studies observed relevance criteria that might be *indirectly* influenced by a user's level of domain knowledge, such as the information depth/scope [4, 21, 29, 31–33], level of detail [7], text quantity [29, 31, 32], and understandability [4, 7]. Two studies observed viewpoint as one of the relevance criteria mentioned by participants [7, 32]. Other studies observed relevance criteria that might be *indirectly* influenced by the viewpoint(s) expressed in the information item: affectiveness (i.e., the user's emotional response to the information) [25, 28] and subjective accuracy (i.e., the user's subjective judgment about the information's accuracy) [4]. Two studies observed relevance criteria associated with experiential information: Schamber et al. [27] observed that some participants preferred information from someone with first-hand knowledge, and Cool et al. [7] observed that some participants preferred interview data. Other studies observed relevance criteria that might be *indirectly* influenced by the author of the information having first-hand experience, such as the information's authority [29], reliability [33], validity [25], and/or credibility [32]. In terms of the geographical dimension, Schamber et al. [27] observed that some participants preferred information associated with a particular location. Finally, in most studies, participants mentioned temporal focus as an important relevance criterion, for example, by stating a preference for recent or historic information [4, 7, 21, 22, 25, 27–29, 31, 32].

Based on these previous findings, we believe that we focused on six extra-topical dimensions that are important to users.

2.3 The Influence of Relevance Criteria on Users' Information Requests

As discussed in the previous section, a large body of research has focused on understanding the wide range of relevance criteria that influence users' relevance decisions. Less research has focused on understanding how specific extra-topical dimensions of an information need might influence the way people request information from a search system.

Kato et al. [16, 17] conducted a large-scale user study that investigated six extra-topical dimensions (referred to as *cognitive search intents* or CSIs) focusing on different attributes of the desired information: (1) exhaustiveness, (2) comprehensibility, (3) objectivity, (4) subjectivity, (5) concreteness, and (6) abstractness. Participants were given a search task and two groups of example documents: a relevant and nonrelevant group with respect to a specific CSI. Participants were asked to produce keyword queries that might return information similar to the relevant examples and dissimilar from the nonrelevant examples. The study found three important trends. First, about half of all queries did *not* have any query terms associated with the assigned CSI. Interestingly, these were cases in which participants *recognized* the CSI in the relevant/nonrelevant examples but *ignored* the CSI when formulating the query. Second, in cases where participants *did* express

the CSI in a query, they did so *indirectly*. Instead of using query terms such as “exhaustive,” “comprehensive,” or “subjective,” participants included query terms associated with the desired online genre (e.g., “blogs”), online source (e.g., Wikipedia), discourse type (e.g., “real experiences”), and/or target user group (e.g., “high school biology”). Finally, when issued to a commercial search engine, queries that included CSI-related language often produced poor search results. Specifically, for three of the six CSIs investigated, queries without any CSI-related language outperformed queries with CSI-related language. The results from Kato et al. [16, 17] suggest that search engines have difficulty processing CSI-related language to predict *extra-topical* relevance, and, perhaps for this reason, users often purposely ignore CSIs when formulating queries.

Joho et al. [15] conducted a user study to investigate search and querying behavior during search tasks with a temporal constraint—the user needs past, recent, or future information. Results found three important trends. First, participants reported higher levels of satisfaction and performance during recent tasks as compared to past and future tasks. Second, the most common strategy adopted by participants was to include temporal expressions in their queries. Participants used different types of temporal expressions, including dates (e.g., “1990”), temporal ranges (e.g., “1990s”), relative expressions (e.g., “after 2014”), temporal categories (e.g., “history”), adjectives (e.g., “latest”), and implicit temporal markers (e.g., “predictions”). Finally, in many cases, temporal expressions failed to produce relevant results. For example, an adjective such as “latest” may not appear in recently published documents, or worse, may appear in old documents. The results from Joho et al. [15] suggests that there is often a vocabulary mismatch between the temporal expressions used in queries and the language present in relevant documents.

We extend this prior work by considering different extra-topical dimensions (or CSIs as referred to in Kato et al. [16, 17]). Additionally, we consider the influence of the medium and target condition on the different strategies adopted by participants when formulating requests.

3 METHODOLOGY

3.1 Study Overview

To explore our three research questions, we gathered a large set of information requests using Amazon Mechanical Turk (MTurk). As described further in Section 3.2, we created 55 search tasks across our six extra-topical dimensions: (1) domain knowledge (10 tasks), (2) viewpoint (10 tasks), (3) experiential (10 tasks), (4) venue location (five tasks), (5) source location (five tasks), and (6) temporal (15 tasks). In addition to manipulating the extra-topical dimension associated with the search task, we manipulated the *medium* of the request (text vs. voice) and the *target* of the request (search engine vs. human intermediary) (Section 3.3). Medium and target were both between-subjects factors. That is, upon accepting their first Human Intelligence Task (HIT), participants were randomly assigned to *one* combination of medium and target condition. The extra-topical dimension was a within-subjects factor. That is, participants were randomly assigned search tasks from all six extra-topical dimensions.

Our HITs were implemented as *external* HITs using our own system in order to control the assignment of experimental units to participants. Participants were allowed to complete as few HITs as desired. However, participants were not assigned the same search task more than once, and could therefore complete a maximum of 55 of our HITs. Each HIT corresponded to one search task and information request, and participants were compensated with \$0.20 USD per HIT.¹ Finally, to help ensure English language proficiency and quality control, we restricted our HITs to U.S. MTurk

¹The study was reviewed and approved by our university’s Institutional Review Board.

workers with a 95% acceptance rate or greater. Ultimately, we gathered information requests from 307 different workers.

Each MTurk HIT proceeded as follows. First, participants were given a set of instructions describing the HIT. The instructions were different depending on the medium and target condition assigned to the participant (Section 3.3). Following these instructions, participants were directed to a webpage where they could access the search task description and submit their information request. We wanted to discourage participants from using parts of the search task description verbatim in their information requests. To this end, participants were required to click a “view task” button that displayed the search task description in a pop-up browser window. Participants had to close the pop-up window to continue interacting with the page, and were therefore not able to see the search task description and produce their request at the same time. Finally, after submitting their information request, participants were directed to a posttask questionnaire (Section 3.4).

We gathered 15 information requests from different MTurk workers for each combination of search task, medium condition, and target condition, for a total of 3,300 requests (i.e., 55 search tasks $\times$ 2 medium conditions $\times$ 2 target conditions $\times$ 15 redundant HITs per experimental condition = 3,300 requests). As described further in Section 3.6, two of the authors manually reviewed all information requests and 39 voice requests were filtered due to noisy input. To maintain a balanced dataset, we randomly sampled 14 information requests per experimental condition to include in our analysis, for a total of 3,080 requests.²

3.2 Extra-Topical Dimensions

We designed 55 search tasks across six extra-topical dimensions: (1) domain knowledge, (2) viewpoint, (3) experiential, (4) venue location, (5) source location, and (6) temporal. All search tasks included a background story that motivated the extra-topical dimension. One of the goals was to avoid influencing participants to use specific vocabulary from the task description in their own information requests (i.e., language priming). To this end, most of our search tasks had the following structure:

“You want to find information on a particular topic to accomplish some task. You have already done some searching, but have only found information that is not appropriate for the intended use (due to extra-topical dimension in question). Now you want to find information that is more appropriate.”

In other words, our search task descriptions described the *opposite* aspect of the corresponding extra-topical dimension, and then instructed the participant to look for information *not* like that.

Next, we define each extra-topical dimension and provide example search tasks.

Domain Knowledge: In our domain knowledge tasks, we simulate a scenario where a user is looking for information on a particular topic but wants information that is appropriate for a domain novice or expert. We created 10 domain knowledge tasks (five novice and five expert tasks) that had the following structure: “You want to find information *to help someone else*. So far, you have only found information that is appropriate for a different audience. You now want to find information that is more appropriate for the person you are trying to help.”

Investigating the influence of domain knowledge on search behavior is not trivial and requires careful design [30]. It is impossible to ask a novice to behave like an expert or vice versa. To circumvent this issue, our domain knowledge manipulation involved situating the task in a scenario where the participant is searching for information on behalf of someone else (i.e., a novice or expert).

²All information requests, along with the corresponding search task description and qualitative annotations, are publicly available here: <https://ils.unc.edu/~jarguell/tois2018/>.

- *Novice Example:* “For his American History class at school, your 13-year-old nephew needs to give a presentation about the health effects of Agent Orange exposure on Vietnam War veterans. You want to help your nephew find information. So far, you have ONLY found highly technical articles intended for historians and medical professionals. You want to find information that is more suitable for your nephew.”
- *Expert Example:* “Your 24-year-old cousin is studying to become an occupational therapist. For one of her college courses, she is writing a research paper on the different health effects of UV rays on the eyes. You want to help her find information for his paper. So far, you have ONLY found information for people who want quick tips on how to protect their eyes and avoid exposure. You want to find information that is more suitable for your cousin.”

Viewpoint: In our viewpoint tasks, we simulate a scenario where someone is looking for information on a particular topic or debate but wants information supporting a particular viewpoint, argument, or stance (i.e., supporting one side of the debate). We created 10 viewpoint search tasks that had the following structure: “You want to gain a more complete understanding of a particular topic or debate. So far, you have only found information supporting one particular viewpoint. You now want to find information that helps you gain a more complete understanding.”

- *Example:* “You are writing a research paper about the history of the European Union (EU) and the Euro. You want to learn more about how the Euro became the common currency among EU members. So far, you have ONLY found historical information about countries, organizations, and individuals that strongly supported the Euro’s adoption. You want to find information that can help you understand the other side of the Euro debate.”

Experiential: In our experiential tasks, we simulate a scenario where someone is looking for information on a particular topic but wants information from someone with first-hand experience. For example, a user might be looking for information about the side effects of a particular medication but may want anecdotal information from patients with first-hand experience, rather than official information from health organizations or drug companies. We created 10 experiential search tasks that had the following structure: “You want to learn about a particular topic. So far, you have only found ‘official’ information originating from organizations or companies that are far removed from ‘real people.’ You now want to find information from people with first-hand experience who can provide anecdotal evidence and advice.”

- *Example:* “Your friend’s grandmother was recently diagnosed with melanoma. You want to help your friend gather information about different treatment plans for melanoma and their effectiveness and side effects. So far, you have ONLY found information published by hospitals and drug companies promoting different treatments. You want to find information from real people about their experiences with different treatment options for melanoma.”

Venue Location: In our venue location tasks, we simulate a scenario where someone is looking for a type of venue in a specific location. We created five venue location search tasks that had the following structure: “You are (or will be) in a particular location conducting some activity. To this end, you want to find a particular type of venue in this location.”

- *Example:* “You are currently on a family trip in New York City. Tomorrow, you plan to visit Central Park and would like to try an NY style pizza restaurant for lunch. Find an NY style pizza restaurant.”

Source Location: In our source location tasks, we simulate a scenario where someone is looking for information on a particular topic but wants to learn about the perspective(s) of people or

sources from a specific location. We created five source location search tasks that had the following structure: "You want to learn more about a particular topic. So far, you have only found information originating from one particular location. Now you want to find information that helps you gain a more complete understanding."

- *Example:* "The U.S. men's soccer team came in fourth place in the 2016 Copa America championship. Personally, you thought this was a great accomplishment and were surprised to hear mostly negative commentary from U.S. news sources. You are interested in learning whether news agencies in other countries also had a negative view of the U.S. soccer team's performance."

Temporal: In our temporal tasks, we simulate a scenario where someone is looking for information on a particular topic but wants information from a specific time period (past, recent, or future). We created 15 temporal search tasks (five past, five recent, and five future tasks) that had the following structure: "You want information about a particular topic. So far, you have only found information from a particular time frame (e.g., only historic information). You now want to find information that helps you gain a more complete understanding."

- *Past Example:* "For her Health class at school, your 15-year-old niece is writing a research paper about the rise of mental illness medications in the past decade. You want to help her find information. So far, you have ONLY found information about mental illness medications currently available on the market. You want to find information that would help your niece gain a more complete picture of the rise of mental illness medications."
- *Recent Example:* "Your nephew is writing a paper for his high school science class about all the different scientific achievements made possible by the Hubble Space Telescope. You want to help him find information for his paper. So far, you have ONLY found information about scientific achievements made shortly after the telescope's launch in 1990. You want to find information that complements these findings."
- *Future Example:* "Your family gave you a telescope for your last birthday and you still haven't used it. You want to find information about cosmic events such as meteor showers that might be worth exploring with a telescope near where you live. So far, you have ONLY found information about cosmic events that have already passed. You want to find information that helps you plan a family outing with your new telescope."

3.3 Medium and Target Conditions

We investigate the influence of two factors on how people request information: the *medium* of the request (text vs. speech) and the *target* of the request (search engine vs. human intermediary). As previously mentioned, participants were assigned to one combination of medium and target condition. Figures 1(a) to 1(d) show all four data collection interfaces. We tried to keep all four data collection interfaces as similar as possible.

Medium: In the *text* condition (Figures 1(a) and 1(c)), participants entered their requests using a textbox of the same size as most commercial search engines. In the *voice* condition (Figures 1(b) and 1(d)), participants were provided with Javascript tools to record their request using their own microphone, save the recording as a WAV file, and upload the WAV file to our server.

Target: In the *search engine* condition (Figures 1(a) and 1(b)), participants were asked: "How would you request this information from a [speech-enabled] search engine such as Google, Bing, or Yahoo?" In the *human intermediary* condition (Figures 1(c) and 1(d)), participants were told that the goal of the HIT was to pilot-test a service called SearchForMe, which allows users to submit

Fig. 1. Data collection interfaces for all medium and target conditions.

information requests to human intermediaries who can search on their behalf. Participants were asked: “How would you request this information from a SearchForMe intermediary?”

3.4 Posttask Questionnaire

Across all medium and target conditions, after submitting their information request, participants completed a posttask questionnaire. Participants were asked about (1) level of difficulty in producing the information request, (2) level of satisfaction with the information request, (3) level of confidence that a search engine (or human intermediary, depending on the target condition) would be able to “understand” the information request, and (4) level of confidence that a search engine (or human intermediary) would be able to find useful information in response to the information request. All four questions were asked using agreement statements and participants responded using a 7-point scale, from strongly disagree (1) to strongly agree (7).

3.5 Voice Request Transcriptions and Relevance Judgments

We used MTurk for two additional data collection efforts. First, all information requests submitted in the voice condition (submitted as WAV files) were transcribed using MTurk. Initially, we tried using several speech-to-text APIs, but the ASR error rate was too high. MTurk workers were paid \$0.20 USD to listen to a WAV file and produce a textual transcription. Participants were allowed to complete a maximum of 100 of these HITs.

Second, as part of RQ2 and RQ3, we were interested in measuring the retrieval effectiveness of information requests. To this end, we issued all requests (in textual form) to a commercial web

search API and cached the top-five search results.³ Then, we used MTurk to gather relevance judgments for these results. MTurk workers were given the original search task description and were asked to judge a set of webpages as being *useful* or *not useful* for the given search task. We decided to use *useful* and *not useful* (rather than *relevant* and *not relevant*) in our instructions in order to influence participants to consider the extra-topical dimension in their judgments. In other words, we wanted MTurk workers to not only consider *topical* relevance but also the result's appropriateness for the task. MTurk workers were paid \$0.20 USD to judge a bundled set of 16 web results for the same task. Each web result was judged by five redundant workers. Again, participants were allowed to complete a maximum of 100 of these HITs. The Fleiss Kappa agreement between workers was $\kappa_f = 0.366$, which is considered "fair agreement" [19].

3.6 MTurk Quality Control

We used MTurk to gather information requests, voice request transcriptions, and relevance judgments for the top-five search results returned by a commercial web search API in response to each request. We implemented several quality control measures during each data collection effort. First, all voice request transcriptions were manually inspected by two of the authors, and 39 voice requests were discarded due to transcription errors. Some transcription errors were due to carelessness and others were due to poor audio quality of the uploaded WAV file (e.g., faint voice, background noise, etc.). In order to maintain a balanced dataset, we then sampled 14 (out of 15) information requests for each combination of search task, medium condition, and target condition, for a total of 3,080 requests that were included in all our analyses (i.e., $55 \times 2 \times 2 \times 14 = 3,080$).

Also, as described above, we gathered relevance judgments using bundled sets of 16 web results associated a given search task. To detect careless MTurk workers, one of these 16 web results originated from a different search task. Workers who marked this extraneous web result as being *useful* were automatically prevented from completing more of our HITs.

3.7 Retrieval Effectiveness

As part of RQ2 and RQ3, we measured the retrieval effectiveness of information requests when issued to a commercial web search API. We measured retrieval effectiveness using NDCG@5 [13]. DCG@5 was computed using the following equation:

$$DCG@5 = \sum_{i=1}^5 \frac{2^{r(i)} - 1}{\log_2(i + 1)},$$

where $r(i)$ denotes the relevance grade of the search result at rank i . Recall that each result was judged by five redundant MTurk workers. We assigned relevance grades to web results based on the fraction of MTurk workers who judged the web result as being *useful*. Finally, NDCG@5 requires computing the DCG@5 value of the *ideal* ranking for the given search task (i.e., the NDCG@5 normalizing factor). We computed the *ideal* DCG@5 value by considering *all* of the top-five search results returned in response to all information requests associated with the given search task (across all medium and target conditions). Using NDCG@5 provided two advantages. First, it allowed us to incorporate the level of agreement between MTurk workers into the evaluation metric. That is, we measured the effectiveness of information requests based on their ability to return results that were *recognized* as being useful by more assessors. Second, as part of RQ3, we compared the retrieval effectiveness of different strategies for requesting information matching a particular extra-topical preference. The normalizing factor in NDCG@5 allowed us to control for the overall

³We decided to focus on the top-five search results based on our budget and the fact that searchers tend to focus their visual attention on the top results [14].

Table 1. Cohen's Kappa (κ_c) Agreement for Codes Associated with Each Extra-Topical Dimension during the Final Round of Qualitative Coding

(a) Domain Knowledge		(b) Viewpoint		(c) Experiential		(d) Venue Location	
Strategy	κ_c	Strategy	κ_c	Strategy	κ_c	Strategy	κ_c
source	1.000	exclude	1.000	source	1.000	proximity	1.000
user	0.978	all sides	0.987	genre	0.985	state	0.973
exclude	0.972	author	0.932	exclude	0.971	city	0.963
purpose	0.958	noun	0.924	discourse	0.951	landmark	0.958
complexity	0.930	proposition	0.915	author	0.915		
genre	0.918						
author	0.798						
(e) Source Location		(f) Temporal					
		Strategy	κ_c			Strategy	κ_c
		adjective	1.000			absolute time point	1.000
		source	1.000			absolute time range	0.988
exclude absolute location			0.985			relative	0.973
absolute location			0.973			verb	0.972
exclude relative location			0.921			all time	0.954
						noun	0.941
						adjective	0.937
						exclude	0.927

difficulty of each search task. In other words, NDC@5 allowed us to focus our RQ3 analysis on the *relative* performance of different strategies for a specific search task.

3.8 Qualitative Coding of Strategies

In our third research question (RQ3), we investigate how our two factors of medium and target influence the way people request information matching a specific type of extra-topical preference. To address this question, for each extra-topical dimension, two of the authors conducted three rounds of qualitative coding. During all three rounds, the authors were not aware of the medium and target condition associated with an information request.

During the first round of coding, both authors independently analyzed information requests associated with a particular extra-topical dimension and used open coding to define a set of strategies reflected in participants' information requests. Then, both authors met to discuss their codes and arrived at a closed set of codes. During the second round of coding, both authors independently coded all requests associated with a particular extra-topical dimension using the closed set of codes. Then, both authors met again to discuss disagreements and potentially refine a particular code's definition if there was high disagreement. During the third and final round of coding, both authors independently recoded all information requests where there was a disagreement after round two.

In the end, an information request was assigned a particular code only if *both* authors agreed on the presence of the strategy in the request. Our codes were intended to be as comprehensive as possible, covering all strategies observed in the data. However, it was still possible for an information request to *not* be assigned any of our codes if the participant did not include any language associated with the extra-topical preference. Furthermore, our codes were applied to information requests independently. Therefore, it was also possible for an information request to be assigned more than one code (i.e., the participant adopted multiple strategies). Tables 1(a) to 1(f) show the

Cohen's Kappa agreement (κ_c) between the two coders for each strategy after the final round of coding. For most strategies, agreement was "almost perfect" ($0.80 < \kappa_c \leq 1.00$) [19].

Next, we describe our closed set of codes for each extra-topical dimension and provide some example phrases associated with each code. Additionally, we provide examples of information requests that did not contain any extra-topical language (referred to as "none") and information requests in which the participant adopted multiple strategies (referred to as "multiple").

3.9 Domain Knowledge Codes

For our domain knowledge dimension, we identified seven codes.

- (1) **Genre:** The participant requested information of a particular genre. Examples: "study guide," "teen guide," "beginner's guide," "mainstream articles," "scientific study," "research paper," "academic journal," "scholarly articles."
- (2) **Complexity:** The participant used adjectives to describe the complexity of the desired information. Examples: "simple," "easy to understand," "brief explanation," "basic," "in-depth," "detailed," "technical."
- (3) **Source:** The participant requested information from a particular online publisher or source. Examples: "for dummies," "wikipedia," "youtube," and "u.s. census."
- (4) **Purpose:** The participant described the purpose of the desired information. Examples: "for a middle-school class project," "for a high school report," "for a class project," "for a college paper," "for a research paper," "for college research."
- (5) **User:** The participant described attributes of the person who would ultimately use the information. Examples: "for kids," "for a beginner," "middle school-er," "for someone studying to become a nutritionist," "for medical professionals," "for occupational therapists."
- (6) **Author:** The participant described attributes of the author of the information. Examples: "from scientists," "from the perspective of a nutritionist," "what does the medical field say."
- (7) **Exclude:** The participant explicitly described *unwanted* information that should be excluded. Examples: "nontechnical articles," "not too technical," "rather than highly technical," "rather than general stuff," "i have already found opinion pieces."
- (8) **None:** The participant did not request information for a domain novice or expert. Examples: "health effects on the eyes from uv rays," "reasons for low voter turnout last election," "long-term effects of using a computer."
- (9) **Multiple:** The participant adopted more than one of the strategies above (1–6). Examples: "computer use research for occupational therapists" (Genre, User), "robotics research only use mainstream articles not journals research articles" (Genre, Exclude), "us presidential elections low voter turnout scientific research not opinion articles" (Genre, Exclude).

3.10 Viewpoint Codes

For our viewpoint dimension, we identified five codes.

- (1) **Noun:** The participant included noun phrases associated with positive or negative argumentation. Examples: "pros/cons," "downsides," "arguments for/against," "risks," "advantages/disadvantages," "justifications," "information supporting/refuting."
- (2) **Proposition:** The participant included a statement or proposition supported by the desired viewpoint. Examples: "rap music doesn't lead to violence," "radio wave exposure causes brain cancer," "increased airport security improving safety," "organic food is bad."

- (3) **Author:** The participant included attributes of the author or publisher of the information that suggests a desire for a particular viewpoint. Examples: “opposers,” “supporters,” “advocates,” “believers,” “skeptics,” “people for/against.”
- (4) **All Sides:** The participant explicitly requested information about all sides of the debate in an attempt to get information from the desired viewpoint. Examples: “different views,” “complete information,” “unbiased information,” “full history,” “all the different sides.”
- (5) **Exclude:** The participant explicitly described *unwanted* information that should be excluded. Examples: “not the healthy side,” “i have already found information stating organic foods are healthy,” “I’ve found arguments for why it is good, but I’d like to find [...],” “i can easily find [...], but I want to see [...].”
- (6) **None:** The participant did not request information supporting a viewpoint. Examples: “rap effect on young people,” “effectiveness of alternative medicine,” “home schooling.”
- (7) **Multiple:** The participant adopted more than one of the strategies above (1–5). Examples: “unbiased sites about the pros and cons of organic food” (Noun, All Sides), “studies supporting that radio waves cause brain cancer” (Noun, Proposition), “brazilian government policies protecting the rain forest [...] only found information criticizing [...].” (Proposition, Exclude).

3.11 Experiential Codes

For our experiential dimension, we identified five codes.

- (1) **Genre:** The participant requested a particular genre or type of online content associated with experiential information. Examples: “reviews,” “blog posts,” “discussion forum,” “message board,” “personal blog.”
- (2) **Discourse:** The participant requested a particular type of discourse associated with experiential information. Examples: “testimonials,” “real stories,” “personal experiences,” “personal accounts,” “anecdotes.”
- (3) **Source:** The participant requested information from a particular online source or publisher associated with experiential information. Examples: “Reddit,” “Consumer Reports.”
- (4) **Author:** The participant described attributes of the author of the desired information. Examples: “patient,” “user,” “cancer survivor,” “from people with melanoma,” “from people who have actually experienced [...].”
- (5) **Exclude:** The participant explicitly described *unwanted* information that should be excluded. Examples: “non-sponsored information,” “not clinics,” “not hospitals,” “not professional opinions,” “[...] rather than professional groups,” “[...] exclude drug companies.”
- (6) **None:** The participant did not request experiential information. Examples: “prostate cancer diagnosis,” “homeopathic osteoporosis remedies,” “aspirin long term usage effects.”
- (7) **Multiple:** The participant adopted more than one of the strategies above (1–5). Examples: “[...] sick building stories blogs real people ” (Genre, Author), “[...] forum or firsthand accounts [...]” (Genre, Discourse) , “[...] personal experiences not doctors [...]” (Discourse, Exclude).

3.12 Venue Location Codes

For our venue location dimension, we identified four codes.

- (1) **City:** The participant specified the city name associated with the desired information. Examples: “in Miami,” “in NYC,” “Chapel Hill.”
- (2) **State:** The participant specified the state name associated with the desired information. Examples: “in Miami, Florida,” “Chapel Hill, NC,” “in Seattle, Washington.”

- (3) **Proximity:** The participant expressed a desire for “nearby” information without stating a location name or landmark. Examples: “nearby,” “near me”.
- (4) **Landmark:** The participant expressed a desire for venues close to a well-known landmark. Examples: “in Central Park,” “around Central Park,” “near Central Park.”
- (5) **None:** The participant did not request venue information for a specific location. Examples: “family style cuban restaurants,” “traditional cuban restaurant,” “movie time for the newest star trek movie.”
- (6) **Multiple:** The participant adopted more than one of the strategies above (1–4). Examples: “chapel hill nc movie theaters” (City, State), “baltimore fresh oysters downtown” (City, Landmark), “ny style pizza near central park nyc” (City, State, Landmark).

3.13 Source Location Codes

For our source location dimension, we identified five codes.

- (1) **Absolute Location:** The participant mentioned the location name (i.e., a proper noun) associated with the desired information. Examples: “in Cleveland,” “in the U.S.,” “UK,” “in Canada.”
- (2) **Adjectives:** The participant used adjectives to describe the origin of the desired information. The adjectives used describe regions relative to the events mentioned in the task description or the participant’s location. Examples: “local,” “national,” “international,” “foreign,” “global.”
- (3) **Source:** The participant requested information from a particular online publisher or source. Examples: “BBC,” “New York Times,” “Denver Post,” “Washington Post,”
- (4) **Exclude Absolute Location:** The participant mentioned a location name (i.e., a proper noun) *not* associated with the desired information, which should be excluded. Examples: “non-US,” “outside of North Carolina,” “from news sources that aren’t based in America,” “from other countries not the U.S.,” “outside U.S.”
- (5) **Exclude Relative Location:** The participant used a relative expression (i.e., not a proper noun) to describe locations not associated with the desired information, which should be excluded. We refer to these as *relative* expressions because the location of interest is implied from the context. Examples: “from other nations,” “[...] rest of the United States,” “what does the rest of the World think [...],” “[...] other countries reporting on [...].”
- (6) **None:** The participant did not request information from a particular location. Examples: “news about u.s. soccer team performance,” “information on the law hb2,” “news about police officer killed in dallas.”
- (7) **Multiple:** The participant adopted more than one of the strategies above (1–5). Examples: “[...] international outlets, not local national media [...].” (Adjectives, Exclude Relative Location), “[...] local news cleveland ohio [...].” (Adjectives, Absolute Location), “[...] local newspaper archive cleveland ohio [...].” (Adjective, Absolute Location).

3.14 Temporal Codes

For our temporal dimension, we identified eight codes.

- (1) **Noun:** The participant used nouns to request a type of information associated with the past, present, or future. Examples: “history,” “major cases,” “news,” “status,” “state of,” “projections,” “prospects,” “the future of.”
- (2) **Adjective:** The participant used adjectives to request information from the past, present, or future. Examples: “past,” “previous,” “older,” “recent,” “current,” “future,” “upcoming,” “projected.”

- (3) **Verb:** The participant used verb phrases in the past, present, or future tense. Examples: “has been,” “has had,” “have been done,” “going on,” “is being used,” “is happening,” “is going to do,” “is going to change,” “will occur.”
- (4) **Absolute Time Point:** The participant requested information from an absolute time period. Examples: “in 2015,” “2016,” “2017,” “in like 2030.”
- (5) **Absolute Time Range:** The participant requested information from an absolute time range. Examples: “since 2006,” “prior to 2016,” “from 2006 to 2016,” “since 1960,” “after 2017.”
- (6) **Relative:** The participant requested information from a time period relative to an event, rather than specifying an actual date. Examples: “in the last 50 years,” “in the last decade,” “prior to today,” “in the last five years,” “post Saddam Hussein’s regime,” “after the debate last year,” “following allegations,” “in the next five years.”
- (7) **All Time:** The participant requested information from all time periods in an attempt to find information from the past, present, or future. Examples: “through the years,” “complete picture of,” “timeline,” “all scientific achievements,” “over time,” “long-term achievements,” “calendar.”
- (8) **Exclude:** The participant explicitly described *unwanted* time periods that should be excluded. Examples: “[...] not the ones currently [...],” “[...] already have current info [...],” “[...] i’ve already looked into older cases [...],” “not historical,” “[...] can only find data for past events [...],” “[...] not the ones that already passed.”
- (9) **None:** The participant did not request information from a particular time period. Examples: “child labor in corporations,” “science experiments in antarctica,” “costa rica tourist industry.”
- (10) **Multiple:** The participant adopted more than one of the strategies above (1–8). Examples: “[...] history of scientific projects in antarctic” (Noun, Verb), “scientific research in antarctica previous articles prior to 2015” (Adjective, Absolute Time Range), “mexico automobile industry future projections” (Adjective, Noun).

4 RQ1 RESULTS: PARTICIPANTS’ PERCEPTIONS

In our first research question (RQ1), we investigate the effects of our two factors (medium and target) on participants’ perceptions about their information request. To address this question, we analyzed participants’ responses to our four posttask questions about (1) level of *difficulty* in producing the information request, (2) level of *satisfaction* with the request, (3) level of confidence that a search engine or human intermediary (depending on the target condition) would be able to *understand* the request, and (4) level of confidence that a search engine or intermediary would be able to *find* useful information in response to the request. Figure 2 shows the means of participants’ responses across medium conditions (T = text, V = voice) and target conditions (S = search engine, H = human intermediary).

We performed two-way ANOVAs to analyze the effects of medium and target on each measure. Medium had a significant main effect for “understand” ($F(1, 3079) = 11.751, p = .001, \eta^2 = .004$) and a *marginally* significant main effect for “find” ($F(1, 3079) = 3.394, p = .066, \eta^2 = .001$). In both cases, participants reported higher levels of confidence in the text condition.

Target had a significant main effect for “difficulty” ($F(1, 3079) = 11.476, p = .001, \eta^2 = .004$), “satisfaction” ($F(1, 3079) = 11.036, p = .001, \eta^2 = .004$), “understand” ($F(1, 3079) = 32.314, p = .000, \eta^2 = .010$), and “find” ($F(1, 3079) = 9.110, p = .003, \eta^2 = .003$). Participants reported lower levels of difficulty and higher levels of satisfaction and confidence in the human intermediary condition.

Fig. 2. Mean (95% confidence intervals) of posttask questionnaire responses across medium and target conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H). Responses were recorded using a 7-point scale from strongly disagree (1) to strongly agree (7).

Finally, medium and target had a significant interaction effect for “find” ($F(1, 3079) = 4.252$, $p = .039$, $\eta^2 = .001$). In the human condition, participants reported roughly equal levels of confidence for both medium conditions (i.e., text and voice). However, in the search engine condition, participants reported *higher* levels of confidence in the text versus voice condition.

5 RQ2 RESULTS: INFORMATION REQUEST CHARACTERISTICS

In our second research question (RQ2), we investigate the effects of our two factors (medium and target) on the characteristics of participants' information requests. We address this question from two perspectives. First, we compare the surface-level characteristics of information requests produced under different conditions. Second, we consider the retrieval effectiveness of participants' information requests when issued to a commercial web search API.

5.1 Surface-Level Characteristics

We compared the surface-level characteristics of information requests produced under different conditions using the following measures: number of words, percentage of stopwords, and percentage of different parts of speech (nouns, adjectives, verbs, prepositions, determiners, pronouns, adverbs, WH-words). Moreover, similar to Guy [12], we used the Stanford NLP Toolkit to compute different parse tree characteristics: the parse tree depth (a measure of grammatical complexity) and whether the parse tree of the request has an “S-root” (full-sentence structure) or an “NP-root” (noun-phrase structure). Table 2 shows the mean values of these measures across medium and target conditions.

We conducted two-way ANOVAs to analyze the effects of medium and target on each of the first 11 *real*-valued measures in Table 2. We used logistic regressions to analyze the effects of medium and target on the last two *binary*-valued measures. The last column in Table 2 indicates whether there was a significant main effect of medium (m) or target (t) and/or a significant interaction effect of medium and target (m*t) on each measure.

The results in Table 2 show three important trends. First, medium had a significant main effect for *all* measures. Compared to textual requests, spoken requests were significantly longer, had a greater percentage of stopwords, had fewer nouns and more of other parts of speech (e.g., verbs, prepositions, wh-words), and had more full-sentence grammatical structure (deeper parse trees,

Table 2. Mean (SD) of Surface-Level Measures Across Conditions: Text/Search Engine (T-S), Voice/Search Engine (V-S), Text/Human (T-H), and Voice/Human (V-H)

	T-S	V-S	T-H	V-H	Sig.
words	6.457 (2.489)	7.777 (3.224)	9.342 (5.338)	14.383 (9.853)	m,t,m*t
stopwords	0.228 (0.187)	0.291 (0.174)	0.316 (0.191)	0.416 (0.182)	m,t,m*t
nouns	0.589 (0.208)	0.526 (0.181)	0.507 (0.187)	0.414 (0.170)	m,t,m*t
adj	0.139 (0.138)	0.131 (0.129)	0.129 (0.119)	0.112 (0.102)	m,t
verbs	0.067 (0.100)	0.088 (0.100)	0.103 (0.102)	0.126 (0.093)	m,t
prep	0.104 (0.105)	0.133 (0.097)	0.128 (0.098)	0.137 (0.083)	m,t,m*t
det	0.021 (0.053)	0.039 (0.063)	0.036 (0.057)	0.061 (0.061)	m,t
pro	0.011 (0.040)	0.015 (0.042)	0.011 (0.034)	0.027 (0.047)	m,t,m*t
adv	0.009 (0.037)	0.011 (0.037)	0.011 (0.035)	0.016 (0.036)	m,t
wh	0.015 (0.042)	0.019 (0.045)	0.023 (0.046)	0.026 (0.041)	m,t
depth	5.842 (1.936)	6.896 (2.491)	7.704 (3.410)	10.282 (5.335)	m,t,m*t
s-root	0.240 (0.428)	0.409 (0.492)	0.457 (0.498)	0.602 (0.490)	m,t
np-root	0.594 (0.491)	0.473 (0.500)	0.417 (0.493)	0.255 (0.436)	m,t

more parse trees with an “S-root” and fewer with an “NP-root”). These results are consistent with previous comparisons between voice and textual queries in the mobile search domain [12], as well as between voice and textual queries gathered in a lab setting for the same tasks [8–10].

Second, target also had a significant main effect for *all* measures. Requests intended for a human intermediary were longer, had a greater percentage of stopwords, had fewer nouns and more of other parts of speech, and had more full-sentence grammatical structure.

Finally, we found a significant interaction effect for six measures: number of words percentage of stopwords, percentage of nouns, percentage of prepositions, percentage of pronouns, and the parse tree depth (grammatical complexity). Medium and target had a strong *additive* effect—participants produced more natural language requests in the voice/human (V-H) condition than in the other three conditions.

5.2 Retrieval Performance

As previously mentioned, to analyze the retrieval effectiveness of participants’ information requests, we used MTurk to gather relevance labels for the top-five results returned from a commercial web search API (Section 3.5). Using these judgments, we measured retrieval performance based on NDCG@5 (Section 3.7). Figures 3(a) to 3(g) show the mean NDCG@5 values across medium and target conditions for information requests associated with each extra-topical dimension and for all tasks combined.

For each subset of information requests, we conducted a multiple linear regression to analyze the effects of the medium and target conditions on NDCG@5 retrieval performance. For medium, the text condition was coded as 0 and the voice condition was coded as 1. For target, the search engine condition was coded as 0 and the human condition was coded as 1.

Table 3 shows the output of each linear regression analysis. All regression equations were significant at the $p < .05$ level and explained between 2% and 8% of the variance (R^2). In terms of medium, the negative β values indicate that NDCG@5 performance was significantly *lower* in the voice versus text condition for all subsets of information requests ($p < .05$). Similarly, in terms of target, the negative β values indicate that NDCG@5 performance was significantly lower in the human versus search engine condition for all subsets of information requests ($p < .05$) *except* for venue location requests ($p = .280$).

Fig. 3. Mean (95% confidence intervals) retrieval performance (based on NDCG@5) for information requests across medium and target conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H). Figures 3(a) to 3(f) show results for each extra-topical dimension and Figure 3(g) shows results for all information requests combined.

Table 3. Multiple Linear Regression Output for Tasks Associated with Each Extra-Topical Dimension and for All Tasks Combined

	Model			Medium		Target	
	R^2	F	p -Value	β	p -Value	β	p -Value
domain knowledge	.081	22.805	.000	−0.186	0.000	−0.215	.000
viewpoint	.022	6.263	.002	−0.095	0.024	−0.114	.007
experiential	.048	14.159	.000	−0.155	0.000	−0.157	.000
venue location	.030	4.272	.015	−0.161	0.007	−0.064	.280
source location	.066	9.775	.000	−0.154	0.009	−0.205	.000
temporal	.029	12.338	.000	−0.095	0.006	−0.140	.000
all	.035	55.408	.000	−0.125	0.000	−0.140	.000

6 RQ3 RESULTS: PARTICIPANTS' STRATEGIES

In our third research question (RQ3), we investigate the effects of our two factors (medium and target) on participants' strategies for requesting information matching a specific extra-topical preference. We address RQ3 from three perspectives. First, in Section 6.1, we consider the distribution of strategies across participants' requests irrespective of the medium and target condition. We do this to gain insight into the extent to which participants adopted the strategies that we manually identified and to understand which strategies were common and which were rare. Second, in

Fig. 4. Domain knowledge tasks: distribution of strategies ($n = 560$). We show results for novice tasks ($n = 280$) and expert tasks ($n = 280$).

Section 6.2, we consider the retrieval effectiveness of information requests associated with each strategy. Finally, in Section 6.3, we consider the distribution of strategies across medium and target conditions. Our ultimate goal was to understand which strategies were prominent, and whether participants used more effective strategies in certain medium and target conditions.

Recall that all information requests were coded by two independent coders, and that codes were treated as being independent (i.e., information requests could be associated with zero, one, or more than one code). We considered a strategy as being present in an information request only if *both* coders agreed on its presence (i.e., the most conservative strategy).

6.1 Distribution of Strategies

In this section, we report on the distribution of strategies for each extra-topical dimension. Figures 4 to 9 show the total number of information requests associated with each strategy that we manually identified (Sections 3.9–3.14). Additionally, we report on two strategies derived from our manual annotations. In Figures 4 to 9, the two columns to the right of the dotted line indicate the number of requests that did not employ any of the strategies we identified (“none”) and the number of requests that employed multiple strategies (“multiple”). As previously noted, our manual coding scheme was intended to be comprehensive. Therefore, column “none” indicates the number of requests without any language suggesting a desire from information matching the corresponding extra-topical preference.

Domain Knowledge: Figure 4 shows the number of information requests ($n = 560$) associated with each strategy for requesting information for a novice or expert. We show results for information requests associated with novice search tasks ($n = 280$) and expert search tasks ($n = 280$).

Results show four important trends. First, based on the “none” column, there were 205/560 information requests without any extra-topical language. Interestingly, these were more common for expert search tasks than for novice search tasks (148 vs. 57). One possible explanation is that our novice search tasks were developed around technical topics, influencing participants to *explicitly* request novice-appropriate information in order to avoid the default expert-appropriate information. Second, in many cases (355/560), participants expressed a desire for novice- or expert-appropriate information using at least one of the strategies we identified. Third, certain strategies were more common than others. Common strategies included asking for information of a specific genre (e.g., “scientific articles”), describing the complexity of the desired information (e.g., “simple explanation”), and describing attributes of the user (e.g., “for a middle-school child”). Uncommon strategies included requesting information from a specific source (e.g., “YouTube videos”),

Fig. 5. Viewpoint tasks: distribution of strategies ($n = 560$).

Fig. 6. Experiential tasks: distribution of strategies ($n = 560$).

describing the purpose of the information (e.g., “for a college-level paper”), describing attributes of the author of the information (e.g., “from the perspective of a nutritionist”), and describing unwanted information that should be excluded (e.g., “not for a layperson”). Finally, certain strategies were much more common for novice versus expert search tasks. For example, describing the complexity of the desired information was more common for novice tasks (e.g., “simple,” “easy to understand,” “high level”), while mentioning the desired genre was more common for expert tasks (e.g., “scientific articles,” “research papers,” “technical reports”).

Viewpoint: Figure 5 shows the number of information requests ($n = 560$) associated with each strategy for requesting information from a particular viewpoint. Based on the “none” column, there were 187/560 information requests without any extra-topical language. In other words, in many cases (383/560), participants requested information from a particular viewpoint by employing at least one of the strategies we identified. The two most common strategies included using noun-phrases to describe positive or negative arguments (e.g., “pros,” “cons,” “opposition,” “negative effects”), as well as including propositions or statements supported by the desired viewpoint (e.g., “rap music no effect on teen violence”). Uncommon strategies included specifying attributes of the author (e.g., “skeptic’s view of alternative medicine”), requesting information on all sides of the debate (e.g., “complete information on organic foods”), and specifying unwanted information (e.g., “information about organic food, not the healthy side”).

Experiential: Figure 6 shows the number of information requests ($n = 560$) associated with each strategy for requesting experiential information (i.e., from someone with first-hand

Fig. 7. Venue location tasks: distribution of strategies ($n = 280$).

Fig. 8. Source location tasks: distribution of strategies ($n = 280$).

experience). Based on the “none” column, there were 115/560 information requests without any extra-topical language. In other words, in many cases (445/560), participants adopted at least one of the strategies we identified. The most common strategy was specifying a type of discourse associated with first-hand information (e.g., “testimony,” “personal accounts,” “stories,” “experiences”). Other slightly less common strategies included specifying the desired genre (e.g., “reviews,” “blogs,” “forum discussions”), specifying attributes of the author (e.g., “from real people who have experience”), and using multiple strategies. The least common strategies included specifying the desired online source (e.g., “Reddit,” “Consumer Reports”) and specifying unwanted information (e.g., “exclude drug companies,” “non-sponsored information,” “not information from companies”).

Venue Location: Figure 7 shows the number of information requests ($n = 280$) associated with each strategy for requesting venue information for a particular location. Based on the “none” column, there were only 8/280 information requests without any extra-topical language. In other words, in many cases (272/280), participants adopted at least one of the strategies we identified. Common strategies included mentioning the city name (e.g., “in Miami”) and the state name (e.g., “in Baltimore, Maryland”). Less common strategies included mentioning an important landmark (e.g., “near Central Park”). The least common strategy was to express a desire for information in close proximity (e.g., “near me,” “nearby”).

Source Location: Figure 8 shows the number of information requests ($n = 280$) associated with each strategy for requesting information from a source originating from a particular

Fig. 9. Temporal tasks: distribution of strategies ($n = 840$). We show results for tasks associated with a temporal preference for past ($n = 280$), recent ($n = 280$), and future information ($n = 280$).

location (e.g., views of the U.S. men's soccer team from sources outside the United States). Based on the "none" column, there were 41/280 information requests without any extra-topical language. In other words, in many cases (239/280), participants adopted at least one of the strategies we identified. The most common strategy was to use adjectives to specify the region of interest (e.g., "international," "foreign," "global," "local," "national"). Slightly less common strategies included specifying a location name associated with the desired information (e.g., "Cleveland newspaper," "U.S. news," "U.K. sports news"), as well as specifying a location name *not* associated with the desired information (e.g., "outside of North Carolina," "not U.S.," "not media in the United States"). Uncommon strategies included specifying the online source or publisher (e.g., "BBC," "USA Today"), using relative expressions to refer to locations that should be excluded (e.g., "from other countries," "the rest of the world"), and using multiple strategies.

Temporal: Figure 9 shows the number of information requests ($n = 840$) associated with each strategy for requesting information from a particular time period. We show results for information requests associated with past ($n = 280$), recent ($n = 280$), and future ($n = 280$) information.

Results show four important trends. First, based on the "none" column, there were 178/840 information requests without any extra-topical language, and these requests were roughly equally common for search tasks requiring past information (52/280), recent information (52/280), and future information (74/280). Second, in many cases (662/840), participants adopted at least one of the strategies we identified. Third, certain strategies were more common than others. Common strategies included using noun phrases to describe information from a particular time period (e.g., "history," "status," "projections"); using adjectives to describe the desired time period (e.g., "past scientific projects," "recent cases," "upcoming cosmic events"); using past, present, and future tense verb phrases (e.g., "have occurred," "is being used," "is going to do"); using relative temporal expressions (e.g., "in the last fifty years," "in the present," "in the near future"); and adopting multiple strategies. Uncommon strategies included specifying an absolute period in time (e.g., "in 2016," "in the year 2030"), specifying an absolute time range (e.g., "prior to 2016," "after 2005," "after 2017"), requesting information from *all* time periods (e.g., "timeline," "over the years," "throughout history"), and specifying unwanted information (e.g., "not the ones currently available," "not historical," "not the ones that have already passed"). Finally, some strategies were more common than others when requesting past, recent, and future information. For example, noun phrases were more common for past and future tasks (e.g., "history," "projections") than for recent tasks; adjectives

Fig. 10. Domain knowledge: NDCG@5 retrieval performance across strategies.

were more common for recent and future tasks (e.g., “current,” “upcoming”) than for past tasks; and specifying an absolute time range was more common for recent tasks (e.g., “2017”) than for past and future tasks.

6.2 Retrieval Performance across Strategies

In this section, we examine the retrieval performance (based on NDCG@5) of different strategies associated with each extra-topical dimension. Figures 10 to 15 show average NDCG@5 performance for each strategy. Strategies are shown in descending order of NDCG@5 performance. The gray columns represent strategies that we explicitly coded for. Conversely, the black columns (“none” and “multiple”) represent strategies that we did not explicitly code for. The “none” column corresponds to the average performance of requests that did not employ any of our strategies (i.e., the participant did not include any extra-topical language in the request). The “multiple” column corresponds to the average performance of requests that employed more than one of our strategies.

To analyze the influence of each strategy on NDCG@5 performance, for each extra-topical dimension, we conducted a multiple linear regression to predict an information request’s NDCG@5 value based on the strategies present in the request. In all cases, the absence and presence of a strategy were coded as 0 and 1, respectively. We excluded “none” and “multiple” from each analysis because these two codes are correlated with the others—“none” is represented by all 0s and “multiple” is represented by multiple 1s. The symbols ▲ and ▼ in Figures 10 to 15 denote a significant increase and decrease in NDCG@5 performance for a specific strategy, respectively.

Domain Knowledge: Figure 10 shows NDCG@5 performance for different strategies for requesting information for a novice or expert. A significant multiple linear regression equation was found ($F(7, 553) = 30.888, p = .000$), with an R^2 of .296. Six strategies had a significant *negative* effect on NDCG@5 performance: “genre” ($\beta = -.127, p = .001$), “user” ($\beta = -.389, p = .000$), “complexity” ($\beta = -.275, p = .000$), “purpose” ($\beta = -.193, p = .000$), “author” ($\beta = -.119, p = .001$), and “exclude” ($\beta = -.267, p = .000$). The best-performing strategy was to not explicitly request novice- or expert-appropriate information using any of our strategies (i.e., “none”), while the worst-performing strategy was to describe unwanted information (i.e., “exclude”).

Viewpoint: Figure 11 shows NDCG@5 performance across different strategies for requesting information from a specific viewpoint. A significant regression equation was found ($F(5, 555) = 16.388, p = .000$), with an R^2 of .129. In this case, “noun” had a significant *positive* effect on NDCG@5 ($\beta = .262, p = .000$). Conversely, two strategies had a significant *negative* effect on NDCG@5: “author” ($\beta = -.141, p = .001$) and “exclude” ($\beta = -.190, p = .000$).

Experiential: Figure 12 shows NDCG@5 performance across different strategies for requesting experiential information. A significant multiple linear regression equation was found ($F(5, 555) = 9.056, p = .000$), with an R^2 of .076. Two strategies had a significant *positive* effect on NDCG@5:

Fig. 11. Viewpoint: NDCG@5 retrieval performance across strategies.

Fig. 12. Experiential: NDCG@5 retrieval performance across strategies.

Fig. 13. Venue location: NDCG@5 retrieval performance across strategies.

“genre” ($\beta = .201, p = .000$) and “discourse” ($\beta = .126, p = .006$). Conversely, two strategies had a significant *negative* effect on NDCG@5: “author” ($\beta = -.128, p = .002$) and “exclude” ($\beta = -.143, p = .001$).

Venue Location: Figure 13 shows NDCG@5 performance across different strategies for requesting venue information from a specific location. A significant multiple linear regression equation was found ($F(4, 277) = 7.226, p = .000$), with an R^2 of .095. One strategy had a significant *negative* effect on NDCG@5: “proximity” ($\beta = -.244, p = .000$).

Source Location: Figure 14 shows NDCG@5 performance across different strategies for requesting information from a specific geographical perspective. A significant multiple linear regression equation was found ($F(5, 275) = 6.493, p = .000$), with an R^2 of .106. One strategy had a significant *positive* effect on NDCG@5: “absolute location” ($\beta = .199, p = .002$). Conversely, two strategies had a significant *negative* effect on NDCG@5: “adjective” ($\beta = -.186, p = .011$) and “exclude relative location” ($\beta = -.133, p = .038$).

Fig. 14. Source location: NDCG@5 retrieval performance across strategies.

Fig. 15. Temporal: NDCG@5 retrieval performance across strategies.

Fig. 16. Domain knowledge: percentage of requests employing different strategies across medium and target conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H). Symbols “medium*” and “target*” denote a significant effect of medium and target on the presence a strategy.

Temporal: Figure 15 shows NDCG@5 performance across different strategies for requesting information from a particular time period. A significant multiple linear regression equation was found ($F(8, 832) = 10.077, p = .000$), with an R^2 of .089. Five strategies had a significant *negative* effect on NDCG@5: “adjective” ($\beta = -.106, p = .003$), “all time” ($\beta = -.119, p = .001$), “relative” ($\beta = -.229, p = .000$), “absolute time point” ($\beta = -.108, p = .001$), and “exclude” ($\beta = -.134, p = .000$).

6.3 Strategies across Medium and Target Conditions

In this section, we examine the distribution of strategies across medium and target conditions for different extra-topical dimensions. Figures 16 to 21 show the percentage of information requests associated with each strategy across medium and target conditions for each extra-topical

Fig. 17. Viewpoint: percentage of requests employing different strategies across medium and target conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H). Symbols “medium*” and “target*” denote a significant effect of medium and target on the presence a strategy.

dimension, respectively. To analyze the influence of the medium and target condition on each strategy, we used a logistic regression to predict the presence/absence of the strategy based on the information request’s medium and target condition. For medium condition, text was coded as 0 and voice was coded as 1. For target condition, search engine was coded as 0 and human was coded as 1.

One of our goals was to determine whether participants used more effective strategies in certain medium and target conditions. To aid with this analysis, in Figures 16 to 21, strategies are ordered from left to right in descending order of NDCG@5 performance (Section 6.2). That is, in each figure, the leftmost strategy was the best-performing one and the rightmost was the worst-performing one.

Domain Knowledge: Figure 16 shows the percentage of requests associated with each strategy for requesting novice- or expert-appropriate information across medium and target conditions. Medium condition was a significant predictor for “complexity” (Wald $\chi^2(1) = 4.284, p = .038$). Participants were more likely to mention the complexity of the desired information in the voice versus text condition (odds ratio = 1.681). Target condition was a significant predictor for “complexity” (Wald $\chi^2(1) = 10.319, p = .001$), “multiple” (Wald $\chi^2(1) = 13.279, p = .000$), and “exclude” (Wald $\chi^2(1) = 5.601, p = .018$), and a *marginally* significant predictor for “purpose” (Wald $\chi^2(1) = 3.653, p = .056$). In the human condition, participants were more likely to mention the complexity of the desired information (odds ratio = 2.276), use multiple strategies (odds ratio = 7.285), describe unwanted information (odds ratio = 4.630), and describe the purpose of the desired information (odds ratio = 2.418).

Viewpoint: Figure 17 shows the percentage of requests associated with each strategy for requesting information from a particular viewpoint across medium and target conditions. Medium was a significant predictor for “exclude” (Wald $\chi^2(1) = 3.958, p = .047$) and a *marginally* significant predictor for “author” (Wald $\chi^2(1) = 3.593, p = .058$). In the voice condition, participants were more likely to describe unwanted information (odds-ratio = 8.364) and to describe attributes of the author (odds ratio = 2.192). Target was a significant predictor for “all sides” (Wald $\chi^2(1) = 3.927, p = .048$) and “multiple” (Wald $\chi^2(1) = 10.191, p = .001$). In the human condition, participants were more likely to request information about all viewpoints (odds ratio = 2.013) and to use multiple strategies (odds ratio = 3.169).

Experiential: Figure 18 shows the percentage of requests associated with each strategy for requesting experiential information across medium and target conditions. Medium condition was not a significant predictor for any strategy. Conversely, target condition was a significant

Fig. 18. Experiential: percentage of requests employing different strategies across medium and target conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H). Symbol “target*” denotes a significant effect of target on the presence a strategy.

Fig. 19. Venue location: percentage of requests employing different strategies across medium and target conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H).

predictor for “multiple” (Wald $\chi^2(1) = 11.801, p = .001$), “author” (Wald $\chi^2(1) = 15.507, p = .000$), and “exclude” (Wald $\chi^2(1) = 7.150, p = .007$). In the human condition, participants were more likely to use multiple strategies (odds ratio = 2.052), to describe attributes of the author (odds ratio = 2.249), and to describe unwanted information (odds ratio = 16.002).

Venue Location: Figure 19 shows the percentage of requests associated with each strategy for requesting venue information from a particular location across medium and target conditions. Neither medium or target condition was a significant predictor for any strategy.

Source Location: Figure 20 shows the percentage of requests associated with each strategy for requesting information from a particular geographical source across medium and target conditions. Medium condition was not a significant predictor for any strategy. Conversely, target was a significant predictor for “multiple” (Wald $\chi^2(1) = 5.647, p = .017$) and “exclude absolute location” (Wald $\chi^2(1) = 7.321, p = .007$). In the human condition, participants were more likely to use multiple strategies (odds ratio = 4.712) and to explicitly mention unwanted location names (odds ratio = 3.052).

Temporal: Figure 21 shows the percentage of requests associated with each strategy for requesting information from a particular time period across medium and target conditions. Medium condition was not a significant predictor for any strategy. Conversely, target condition was a significant predictor for “verb” (Wald $\chi^2(1) = 6.941, p = .008$), “relative” (Wald $\chi^2(1) = 8.477, p = .004$),

Fig. 20. Source location: percentage of requests employing different strategies across medium and target conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H). Symbol “target*” denotes a significant effect of target on the presence a strategy.

Fig. 21. Temporal: percentage of requests employing different strategies across medium and target conditions: text/search engine (T-S), voice/search engine (V-S), text/human (T-H), and voice/human (V-H). Symbol “target*” denotes a significant effect of target on the presence a strategy.

“multiple” (Wald $\chi^2(1) = 21.204$, $p = .000$), and “exclude” (Wald $\chi^2(1) = 5.703$, $p = .017$). In the human condition, participants were more likely to use temporal verbs (odds ratio = 1.704), use relative temporal expressions (odds ratio = 1.705), use multiple strategies (odds ratio = 2.329), and explicitly mention unwanted time periods (odds ratio = 12.086).

7 DISCUSSION

7.1 Research Question 1

In terms of RQ1 (Section 4), our results suggest three main trends. First, target had a greater effect than medium on participants' perceptions about their own requests (Figure 2). Participants reported greater ease, satisfaction, and confidence when producing requests for a human intermediary versus a search engine. In other words, participants had a better experience formulating requests when they had raised expectations about the “system's” ability to understand and respond to requests.

Second, medium and target had a significant interaction effect for one of our confidence questions (i.e., “find”). When producing requests for a search engine, participants reported *lower* levels of confidence in the voice versus text condition. It may be that participants feared the possibility of

speech recognition errors. Alternatively, perhaps some of this lower confidence can be attributed to the inherent nature of voice queries. Compared to textual requests, voice requests cannot be constructed incrementally and edited (e.g., by deleting terms) before being submitted to a system.

Finally, while some of the differences were significant, they were also small (considering a 7-point scale and the small effect sizes). It is worth noting that we used a between-subjects design (i.e., participants were assigned to one combination of medium and target condition). We might have observed larger differences if individual participants had experienced different target and medium conditions (by grounding their responses on other conditions).

7.2 Research Question 2

In terms of RQ2 (Section 5), our results suggest two main trends. First, both of our factors influenced the structure of information requests. Consistent with prior work, voice requests were longer and had grammatical structures closer to natural language than textual requests [12]. We observed the same trends for requests intended for a human intermediary versus a search engine. Furthermore, we found a strong *additive* effect of both conditions. That is, voice queries intended for a human were the longest and had the greatest resemblance to natural language (e.g., had the deepest parse trees and most often had parse trees with full-sentence structure).

Second, both factors also influenced retrieval performance (based on NDCG@5). Consistent with prior work, unmodified textual requests outperformed unmodified voice requests [8]. Because our requests were manually transcribed, the drop in performance for voice requests was not due to speech recognition errors. We observed the same trend for requests intended for a human. Moreover, both factors had a strong *additive* effect—textual requests intended for a search engine performed about 30% better than voice queries intended for a human.

An important follow-up question is: what caused requests to perform poorly in the voice and human intermediary conditions? In the next section (as part of RQ3), we examine whether some of this drop in performance was caused by participants adopting more complex strategies for conveying an extra-topical preference. However, we also wanted to know whether requests produced in the voice and human intermediary conditions were more likely to contain terms that are not related to the information need in any way (topical or extra-topical).

To help answer this question, we aimed to find the most distinctive terms in voice requests (compared to textual requests) and the most distinctive terms in requests intended for a human (compared to search engine requests). Table 4 lists the top 20 terms with the greatest contribution to the KL-divergence (KLD) between four pairs of language models: (1) textual versus voice requests, (2) voice versus textual requests, (3) search engine versus human requests, and (4) human versus search engine requests. A term's contribution to the KLD between language models θ_1 and θ_2 is given by

$$KLD(w, \theta_1, \theta_2) = P(w|\theta_1) \log_2 \left(\frac{P(w|\theta_1)}{P(w|\theta_2)} \right).$$

A term's contribution to the KLD is greater when the term has a *high* probability in the first language model (θ_1) and a *low* probability in the second language model (θ_2). In this respect, the second column in Table 4 shows terms that were relatively more common in voice requests as compared to textual requests, while the fourth column in Table 4 shows terms that were relatively more common in human requests as compared to search engine requests. To aid with the analysis, the gray cells in Table 4 correspond to terms that we thought were not related to the search task in any way (topical or extra-topical).

Table 4 shows two important trends. First, by comparing the second column with the first, it is evident that voice requests were more likely than textual requests to contain terms that were not

Table 4. Top 20 Terms with Greatest Contribution to the KL-Divergence between Language Models Generated from Textual versus Voice Requests (Column 1), Voice versus Textual Requests (Column 2), Search Engine versus Human Requests (Column 3), and Human versus Search Engine Requests (Column 4)

Text vs. Voice	Voice vs. Text	Search Engine vs. Human	Human vs. Search Engine
nc	find	hey	i
looking	the	google	the
kids	me	experiences	like
history	search	history	to
advance	to	experience	that
ny	i'd	research	information
stories	information	personal	about
am	like	us	please
2016	about	with	would
thanking	show	near	and
experiences	that	effects	or
cancer	i	air	i'd
theft	a	of	find
future	want	kids	a
use	or	food	can
projections	any	recent	you
usa	can	anorexia	any
park	know	kids	want
patient	would	news	looking
personal	hey	mexico	show

The gray cells show terms that are not related to the search task in any way (topical or extra-topical).

related to the search task. Second, by comparing the fourth column with the third, the same trend was true for human versus search engine requests. Based on these results, it is clear that some of the performance drop for requests produced in the voice and human intermediary conditions was caused by the inclusion of extraneous terms unrelated to the information need.

Arguello et al. [2] focused on automatically reducing voice requests in order to improve performance. The authors found that many of the problematic terms that caused voice queries to perform poorly included extraneous terms not related to the information need (e.g., “find,” “information”). Our results support these findings and suggest that requests intended for a human are also more likely to contain extraneous terms.

7.3 Research Question 3

Distribution of Strategies: In terms of the strategies adopted by participants (Section 6.1), we observed two main trends. First, based on our qualitative coding, participants often requested information matching the assigned extra-topical preference using at least one of the strategies identified. In Figures 4 to 9, the number of requests associated with the “none” category (not containing any explicit extra-topical language) was always less than half the information requests. The “none” category was the least frequent for venue location requests (3%) and the most frequent for domain knowledge requests (57%).

Second, certain strategies (i.e., source, genre, author, and exclude) were present in *multiple* extra-topical dimensions. For example, requesting information from a specific online source was a

strategy for domain knowledge tasks (e.g., “YouTube”), experiential tasks (e.g., “Reddit”), and source location tasks (e.g., “BBC”). Likewise, mentioning attributes of the author was a strategy for domain knowledge tasks (e.g., “nutritionist”), viewpoint tasks (e.g., “skeptic’s view”), and experiential tasks (e.g., “from real people with experience”). Finally, explicitly describing undesired information that should be excluded was a strategy observed in all but one extra-topical dimension. These are all common strategies that search engines should be able to handle. However, some seem more problematic than others. For example, algorithms that rely on term matching may be effective at returning content from a specific online source (e.g., “YouTube”), but less effective at excluding content with characteristics mentioned in the query (e.g., “not for a layperson”).

Retrieval Performance Across Strategies: In terms of retrieval performance (Section 6.2), our results clearly show that certain strategies performed better than others. In terms of successful strategies, for viewpoint tasks, including nouns associated with the desired viewpoint (e.g., “pros,” “cons,” “advantages”) significantly improved NDCG@5 performance. Likewise, for experiential tasks, specifying the genre of the desired information (e.g., “blogs”), as well as describing the desired type of discourse (e.g., “testimonials”), significantly improved NDCG@5 performance. Conversely, in terms of unsuccessful strategies, mentioning attributes of the author significantly hurt retrieval performance for all three extra-topical dimensions where it was an observed strategy: domain knowledge, viewpoint, and experiential. Likewise, explicitly mentioning undesired information in the request significantly hurt retrieval performance for all five extra-topical dimensions where it was observed as a strategy: domain knowledge, viewpoint, experiential, source location, and temporal.

A common strategy across all extra-topical dimensions was *not* explicitly including any extra-topical terms in the request (denoted as “none”). The performance of “none” was mixed. For some extra-topical dimensions, it performed relatively well (e.g., domain knowledge, temporal). However, for other dimensions, it performed poorly (e.g., experiential, venue location). Interestingly, “none” performed noticeably worse for experiential tasks, possibly because most web content does not convey experiential (i.e., first-hand) information. This result has implications for information retrieval research—search algorithms may need to be especially effective in handling extra-topical query terms in cases where there are relatively few documents in the collection matching the extra-topical preference.

Strategies across Medium and Target Conditions: We observed two main trends in terms of the effects of medium and target on the strategies adopted by participants (Section 6.3). First, while both factors had significant effects, target had a much stronger effect. Medium had a significant effect for two strategies: “complexity” (domain knowledge) and “exclude” (viewpoint). For both of these, the strategy was significantly more likely in the voice versus text condition. Target had a significant effect for 14 strategies across all extra-topical dimensions. For all 14 strategies, the presence of the strategy was significantly more likely in the human versus search engine condition.

The second important trend is that both factors influenced participants to adopt strategies that resulted in worse retrieval performance. As previously mentioned, the strategies in Figures 16 to 21 are presented in descending order of NDCG@5 performance (from left to right). This trend can be observed by considering the position of a strategy compared to “none.” As previously noted, medium had a significant effect for two strategies, and both strategies performed worse than “none” within their respective extra-topical dimensions. Similarly, target had a significant effect for 14 strategies, and 11 of these performed worse than “none” within their respective extra-topical dimensions. In other words, when producing voice queries for a human, participants were more likely to use strategies that hurt retrieval performance—a better approach would have been to *not* explicitly mention the extra-topical preference in the request in any way.

7.4 Implications and Opportunities for Future Research

Based on our RQ1 results, participants reported the least amount of difficulty while producing requests in the voice and human intermediary conditions (Figure 2). This result is perhaps not surprising. Voice is a natural means to communicate an information need, and participants may have devoted less effort strategizing their requests when they had higher expectations about the target “system’s” ability to understand and respond to requests. This result suggests that there are benefits from systems that can accept more naturalistic forms of input and systems that raise users’ expectations.

Based on our RQ2 results, while participants had a better experience producing requests in the voice and human intermediary conditions, these requests also produced the worst retrieval performance when issued *unmodified* to a commercial web search API. Our results suggest that requests produced in the voice and human intermediary conditions were problematic for two reasons.

First, these requests were more likely to contain extraneous terms that are completely unrelated to the information need in any way (topical or extra-topical) (Table 4). These are terms that could be automatically dropped or down-weighted. Prior research has developed a wide range of methods for algorithmically pruning queries to improve their performance (see review [11]). Even simple approaches such as dropping all terms except for nouns and adjectives have yielded promising results [8].

Second, in the voice and human intermediary conditions, participants were more likely to adopt problematic strategies for describing extra-topical dimensions of the information need. For example, problematic strategies included describing the complexity of the desired information (domain knowledge), describing unwanted information (domain knowledge, viewpoint, experiential, source location, and temporal), and adopting multiple strategies (domain knowledge, viewpoint, source location, and temporal). While the effect of target was stronger medium, these cases were more common in the voice and human intermediary conditions. Based on our RQ3 results, all of these strategies yielded worse retrieval performance than ignoring the extra-topical dimension in the request (“none”).

An important question for future research is: how should systems deal with information requests that describe extra-topical dimensions of the information need? Automatic query reduction seems like a promising approach for dealing with extraneous terms that are completely unrelated to the user’s information need. However, ignoring extra-topical language in queries seems like a missed opportunity. An important direction for future research is to investigate approaches for *leveraging* (rather than ignoring) extra-topical language in queries in order to improve their retrieval performance.

Consistent with prior work [8, 15–17], we hypothesize that extra-topical query terms can be problematic because they may not appear in relevant documents and/or may appear in nonrelevant documents. There are several directions to explore. Possibly the lowest-hanging fruit is to explore document and query expansion methods. In terms of document expansion, it may be possible to harness extra-topical language about documents from external sources such as anchor text or previous queries with clicks. For example, a webpage suitable for domain novices may not contain the query term “simple,” but anchor text pointing to the webpage might. In terms of query expansion, our results found that some strategies for expressing an extra-topical preference performed better than others. For example, for our source location tasks, requesting information from a specific online source (e.g., “BBC”) yielded better retrieval performance than specifying the desired location using adjectives (e.g., “international”). Future work might consider query expansion methods that can recognize an extra-topical preference in a query and augment the query with *effective* extra-topical language. Another direction to explore involves learning-to-rank (LTR)

approaches that can exploit complex feature interactions (e.g., deep learning approaches). Complex LTR models might be able to associate certain query terms (e.g., “latest”) with specific document attribute-values (e.g., a recent publication date). Finally, outside of information retrieval (IR), a large body of research has focused on natural language interfaces for structured databases [23]. These techniques focus on automatically translating natural language requests into highly structured SQL queries. Within the scope of IR, one could imagine translation models that can convert natural language queries into structured queries that score documents using a combination of query-document similarity, but also favor documents with specific meta-data values (possibly inferred in advance). For example, using the Indri query language, one could imagine approaches that can translate the unstructured query “latest hubble space telescope achievements” into the structured query: `#combine(prior(RECENT) hubble space telescope achievements)`.⁴

7.5 Caveats and Possible Limitations of the Study

Our study has two possible limitations that are worth discussing.

One-Shot Information Requests without Interaction: Search is inherently an interactive process. However, in this study, participants formulated information requests and there was no system output, interaction, or reformulation. Prior studies have also gathered and analyzed information requests without interaction. For example, Bailey et al. [3] conducted a crowdsourced study in which participants were asked to produce queries associated with TREC topics to study the variability of query performance versus system performance. Similarly, prior studies have also gathered voice queries for TREC topics in order to compare the retrieval performance of voice queries versus textual queries [1, 2, 8].

While gathering information requests without subsequent interaction is convenient, it also has potential limitations. For the purpose of our study, it is possible that participants contributed *more* details about the search task (and the search task’s extra-topical dimension) than they would in an interactive setting. Additional research is needed to investigate how searchers address extra-topical preferences in an interactive setting. For example, in the desktop search domain, searchers might partially rely on interactive tools (e.g., metadata filters and faceted search tools) to refine their searches based on their extra-topical preferences. Similarly, in a more “conversational” search domain, searchers might rely on follow-up questions/responses from the system to refine their searches. We view the research in this article as a first step. Future research is needed to understand how people address extra-topical preferences given the ability to formulate complex, natural language requests and given access to interactive mechanisms that can be used to refine their searches after the initial request.

Search Task Manipulation: In this study, we focused on information requests associated with a specific type of extra-topical preference. To this end, we contextualized our search tasks in a scenario (or backstory) that motivated a specific type of information (e.g., information for a domain novice vs. an expert). To avoid language priming, most of our search tasks described a scenario in which the participant had already done some searching but had not found information appropriate for the task (e.g., “So far, you have *ONLY* found highly technical articles intended for medical professionals.”). In this way, we aimed to influence participants without priming them with relevant extra-topical terms (e.g., “simple,” “easy to understand”). It is possible that our task manipulation introduced some bias in our results. For example, we observed cases where participants explicitly stated the types of information they did *not* want. Such cases may have been more likely due to our task manipulation.

⁴<http://lemurproject.org/lemur/IndriQueryLanguage.php>.

From a methodological perspective, this is an interesting research question: how can we influence participants to desire certain types of information without priming them to adopt certain strategies? During pilot testing, we considered using the approach from Kato et al. [16, 17]—giving participants topical search tasks and providing them with examples of relevant and nonrelevant documents based on the extra-topical dimension in question (e.g., example documents suitable for a novice and documents suitable for an expert). This approach, however, also has disadvantages—participants might be influenced by the language in the example documents (language priming) and might not necessarily recognize the extra-topical dimension based on the given examples, resulting in loss of data [16, 17]. Future research is needed to compare between different manipulations in order to understand their advantages and disadvantages.

8 CONCLUSION

In this article, we studied how people formulate information requests under different experimental conditions. We manipulated three factors: the medium of the requests (text vs. voice), the target of the request (search engine vs. human intermediary), and the extra-topical dimension associated with the search task. We focused on six extra-topical dimensions: (1) domain knowledge, (2) viewpoint, (3) experiential, (4) venue location, (5) source location, and (6) temporal. With respect to our three research questions, we found the following trends.

RQ1: The target of the request had a stronger effect on participants' perceptions about their own requests. Participants reported greater ease, satisfaction, and confidence in the human intermediary condition. In other words, participants had a better experience formulating requests when they had higher expectations about the "system's" ability to understand and respond to requests.

RQ2: Both factors had a significant effect on the structure of participants' requests. Requests produced using voice and requests intended for a human were longer, had structures closer to natural language, and yielded *lower* retrieval performance. One explanation for the lower performance is that requests produced using voice and intended for a human were more likely to contain terms not central to the information need in any way (topical or extra-topical).

RQ3: Participants used a wide range of strategies for requesting information matching a specific type of extra-topical preference. In terms of retrieval performance, some strategies were more effective than others. For example, requesting information from a specific online source yielded comparatively good performance, while explicitly describing unwanted information that should be excluded from the search results yielded comparatively bad performance. A common strategy was to completely ignore the extra-topical dimension in the request. For two extra-topical dimensions (domain knowledge and temporal), this was the best-performing strategy, suggesting a need for retrieval algorithms that can handle extra-topical preferences described in a user's request.

Both of our factors of medium and target had an influence on the presence/absence of certain strategies. However, target had a stronger effect. Most importantly, almost all of the strategies that were significantly more likely in the human intermediary condition (11 out of 14) had *lower* retrieval performance than requests that completely ignored the extra-topical dimension.

Our results have implications for future information retrieval research. As search interfaces evolve and afford more naturalistic interactions, users may raise their expectations about the system's ability to understand and respond to requests. In turn, people may adjust the way they formulate queries. Users may produce longer queries that more closely resemble natural language. Moreover, people may use more complex strategies for conveying extra-topical criteria associated with the information need. A large body of prior work has focused on automatically *reducing* verbose queries to improve performance [11]. Such methods might be effective at ignoring query terms that are unrelated to the information need. However, dropping or down-weighting extra-topical query terms is a missed opportunity. Future research should focus on developing retrieval

techniques that can harness (rather than ignore) extra-topical language in the query in order to improve performance. In this article, we outlined a few directions to explore.

REFERENCES

- [1] Jaime Arguello, Sandeep Avula, and Fernando Diaz. 2016. Using query performance predictors to improve spoken queries. In *ECIR*. Springer.
- [2] Jaime Arguello, Sandeep Avula, and Fernando Diaz. 2017. Using query performance predictors to reduce spoken queries. In *ECIR*. Springer.
- [3] Peter Bailey, Alistair Moffat, Falk Scholer, and Paul Thomas. 2015. User variability and IR system evaluation. In *SIGIR*. ACM, 625–634.
- [4] Carol L. Barry. 1994. User-defined relevance criteria: An exploratory study. *Journal of the Association for Information Science and Technology* 45, 3 (1994), 149–159.
- [5] Carol L. Barry and Linda Schamber. 1998. Users' criteria for relevance evaluation: A cross-situational comparison. *IP&M* 34, 2–3 (1998), 219–236.
- [6] Judy Bateman. 1998. Changes in relevance criteria: A longitudinal study. In *ASIST*. 23–32.
- [7] Collen Cool, Nick Belkin, Ophir Frieder, and Paul Kantor. 1993. Characteristics of texts affecting relevance judgments. In *Proceedings of the 14th National Online Meeting*. 77–84.
- [8] Fabio Crestani and Heather Du. 2006. Written versus spoken queries: A qualitative and quantitative comparative analysis. *Journal of the Association for Information Science and Technology* 57, 7 (2006), 881–890.
- [9] Heather Du and Fabio Crestani. 2003. Spoken versus written queries for mobile information access. In *MobileHCI*.
- [10] Heather Du and Fabio Crestani. 2005. Spoken versus written queries for mobile information access: An experiment on Mandarin chinese. In *IJCNLP*. Springer, 745–754.
- [11] Manish Gupta and Michael Bendersky. 2015. Information retrieval with verbose queries. *Foundations and Trends in Information Retrieval* 9, 3–4 (2015), 209–354.
- [12] Ido Guy. 2016. Searching by talking: Analysis of voice queries on mobile web search. In *SIGIR*. ACM, 35–44.
- [13] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *Transactions on Information Systems* 20, 4 (Oct. 2002), 422–446.
- [14] Thorsten Joachims, Laura Granka, Bing Pan, Helene Hembrooke, and Geri Gay. 2005. Accurately interpreting click-through data as implicit feedback. In *SIGIR*. ACM, 154–161.
- [15] Hideo Joho, Adam Jatowt, and Roi Blanco. 2015. Temporal information searching behaviour and strategies. *Information Processing & Management* 51, 6 (2015), 834–850.
- [16] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. 2014. Cognitive search intents hidden behind queries: A user study on query formulations. In *WWW*. ACM, 313–314.
- [17] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. 2014. Investigating users' query formulations for cognitive search intents. In *SIGIR*. ACM, 577–586.
- [18] Giridhar Kumaran and Vitor R. Carvalho. 2009. Reducing long queries using query quality predictors. In *SIGIR*. ACM, 564–571.
- [19] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159–174.
- [20] Matthew Lease. 2009. An improved Markov random field model for supporting verbose queries. In *SIGIR*. ACM, 476–483.
- [21] Kelly L. Maglaughlin and Diane H. Sonnenwald. 2002. User perspectives on relevance criteria: A comparison among relevant, partially relevant, and not-relevant judgments. *Journal of the Association for Information Science and Technology* 53, 5 (2002), 327–342.
- [22] Taemin Kim Park. 1993. The nature of relevance in information retrieval: An empirical study. *Library Quarterly* 63, 3 (1993), 318–351.
- [23] Rodolfo A. Pazos R., Juan J. González B., Marco A. Aguirre L., José A. Martínez F., and Héctor J. Fraire H. 2013. *Natural Language Interfaces to Databases: An Analysis of the State of the Art*. Springer, Berlin, 463–480.
- [24] Tefko Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the Association for Information Science and Technology* 58, 13 (2007), 1915–1933.
- [25] Reijo Savolainen and Jarkko Kari. 2006. User's defined relevance criteria in web searching. *Journal of Documentation* 62, 6 (2006), 685–707.
- [26] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. *Your Word Is My Command: Google Search by Voice: A Case Study*. Springer US, 61–90.
- [27] Linda Schamber. 1991. Users' criteria for evaluation in a multimedia environment. In *ASIST*.

- [28] Amanda Spink, Howard Greisdorf, and Judy Bateman. 1998. From highly relevant to not relevant: Examining different regions of relevance. *Information Processing & Management* 34, 5 (1998), 599–621.
- [29] Anastasios Tombros, Ian Ruthven, and Joemon M. Jose. 2005. How users assess web pages for information seeking. *Journal of the Association for Information Science and Technology* 56, 4 (2005), 327–344.
- [30] Barbara M. Wildemuth. 2004. The effects of domain knowledge on search tactic formulation. *Journal of the Association for Information Science and Technology* 55, 3 (2004), 246–258.
- [31] Iris Xie and Edward Benoit. 2013. Search result list evaluation versus document evaluation: Similarities and differences. *Journal of Documentation* 69, 1 (2013), 49–80.
- [32] Iris Xie, Edward Benoit, and Huan Zhang. 2010. How do users evaluate individual documents? An analysis of dimensions of evaluation activities. In *Proceedings of the 7th International Conference on Conceptions of Library and Information Science*.
- [33] Yunjie Xu and Zhiwei Chen. 2006. Relevance judgment: What do information users consider beyond topicality? *Journal of the Association for Information Science and Technology* 57, 7 (2006), 961–973.
- [34] Jeonghe Yi and Farzin Maghoul. 2011. Mobile search pattern evolution: The trend and the impact of voice queries. In *WWW*. ACM, 165–166.

Received November 2017; revised March 2018; accepted April 2018