Tip of the Tongue Known-Item Retrieval

A Case Study in Movie Identification

Jaime Arguello
University of North Carolina at Chapel Hill
jarguell@email.unc.edu

Adam Ferguson
Microsoft
adfergus@microsoft.com

Emery Fine
Microsoft
emfine@microsoft.com

Bhaskar Mitra
Microsoft
bhaskar.mitra@microsoft.com

Hamed Zamani*
University of Massachusetts Amherst
zamani@cs.umass.edu

Fernando Diaz†
Microsoft
diazf@acm.org

ABSTRACT

While current information retrieval systems are effective for known-item retrieval where the searcher provides a precise name or identifier for the item being sought, systems tend to be much less effective for cases where the searcher is unable to express a precise name or identifier. We refer to this as *tip of the tongue (TOT) known-item retrieval*, named after the cognitive state of not being able to retrieve an item from memory. Using movie search as a case study, we explore the characteristics of questions posed by searchers in TOT states in a community question answering website. We analyze how searchers express their information needs during TOT states in the movie domain. Specifically, what information do searchers remember about the item being sought and how do they convey this information? Our results suggest that searchers use a combination of information about: (1) the *content* of the item sought, (2) the *context* in which they previously engaged with the item, and (3) previous attempts to find the item using other resources (e.g., search engines). Additionally, searchers convey information by sometimes expressing uncertainty (i.e., hedging), opinions, emotions, and by performing relative (vs. absolute) comparisons with attributes of the item. As a result of our analysis, we believe that searchers in TOT states may require specialized query understanding methods or document representations. Finally, our preliminary retrieval experiments show the impact of each information type presented in information requests on retrieval performance.

ACM Reference Format:
Jaime Arguello, Adam Ferguson, Emery Fine, Bhaskar Mitra, Hamed Zamani, and Fernando Diaz. 2021. Tip of the Tongue Known-Item Retrieval: A Case Study in Movie Identification. In *Proceedings of the 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '21), March 14–19, 2021, Canberra, ACT, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3406522.3446021

*A part of this work was done while Hamed Zamani was affiliated with Microsoft.
†Now at Google.

CHIIR '21, March 14–19, 2021, Canberra, ACT, Australia

ACM ISBN 978-1-4503-8055-3/21/03…$15.00
https://doi.org/10.1145/3406522.3446021

1 INTRODUCTION

Known-item retrieval refers to a broad class of scenarios where a searcher's information need is for a single, specific document known to exist in the corpus [26]. Searchers may seek an item they have seen before [10] or one they believe exists [4]. During known-item retrieval, searchers may express their need using a unique identifier (e.g., title), bibliographic information (e.g., author or genre), or content cues (e.g., keywords). Information retrieval systems, for example web search engines, can leverage previously issued queries and engagements to improve known-item search, although such techniques are less effective for unpopular or new documents with less behavioral data [9].

In this paper, we investigate known-item retrieval scenarios where a searcher is looking for a previously seen item but is *not* able to express precise or even reliable information about the item. Imprecision may result from a long delay between the information need and when the searcher most recently engaged with the item (i.e., long-term memory degradation) or from the lack of a universally adopted content description language (e.g., searching for a song based on its drum beat or a book based on its narrative structure). While similar to re-finding tasks, the emphasis on identification—as opposed to navigation—makes these information needs acute, one-off episodes—as opposed to repeated requests.

We refer to a searcher in this situation as being in a 'tip of the tongue' (TOT) state, due to its similarity to the cognitive state of not being able to retrieve an item from memory [6]. We adopt the following definition of a TOT information need,

> *an item identification task where the searcher has previously experienced or consumed the item but cannot recall a reliable identifier.*

Elsweiler et al. [12] found that searchers specifically in TOT states tend to be prone to higher levels of frustration compared to other memory lapse states.

In recent years, several community question-answering (CQA) sites have emerged to support searchers with TOT information needs.[1] These sites are tailored for searchers who want to find a specific known item (e.g., a movie, book, song, artist, video game,

[1]https://www.reddit.com/r/tipofmytongue, https://irememberthismovie.com, https://www.watzatsong.com/en, https://www.goodreads.com/group/show/185-what-s-the-name-of-that-book, https://scifi.stackexchange.com/questions/tagged/story-identification

childhood toy) but do not remember its name or other unique metadata. Searchers pose questions, composed of a title and description, while community members ask clarification questions and suggest answers. A thread is typically closed (or "solved") when the questioner indicates that the correct answer has been provided. Importantly, one might view such CQA sites as serving people with TOT information needs that could not be resolved using other resources, *including search systems* [28]. To paraphrase a blogger's comment about `watzatsong.com`, "when computers fail, seek the help of humans."[2]

In order to study TOT information needs, we analyzed questions (i.e., information *requests*) posted to the 'I Remember This Movie' TOT CQA site.[3] We decided to use movie identification as an initial case study because the movie domain is associated with rich auxiliary data available online, including metadata and long-form plot descriptions.

Our research in this paper focuses on two main research questions. First, we are interested in understanding TOT search requests,

> **RQ1**: How does a searcher in a TOT state express their need in the search request?

To answer this question, we investigated the characteristics of TOT information requests. Our goal is to understand the types of phenomena present in TOT requests. Insights gained as part of **RQ1** help us answer two basic questions: (1) What do searchers in TOT states remember? and (2) How do they convey this information? To address **RQ1**, we conducted an extensive qualitative analysis of 1,000 requests posted to 'I Remember This Movie'. Our qualitative coding scheme (applied at the sentence-level) was developed in a bottom-up fashion. After analyzing a subset of the data, we developed qualitative codes along four dimensions: (1) the *target* of the sentence (i.e., what is the sentence about?), (2) the presence of opinions or emotional expressions, (3) expressions of uncertainty (i.e., hedging), and (4) the presence of relative comparisons (versus descriptions in absolute terms). In terms of the *target*, sentences described characteristics of the movie itself, the *context* in which the searcher watched the movie, previous (failed) attempts to find the movie, or neither (i.e., social nicety).

We were also interested in understanding the performance of automatic retrieval for TOT needs,

> **RQ2**: How well does a conventional retrieval system satisfy TOT requests?

To answer this question, we gathered a corpus of movie plot descriptions and evaluated the effectiveness of TOT requests when issued as queries to a standard information retrieval system. We studied the relationship between retrieval performance and the presence/absence of codes developed as part of **RQ1**. For example, does performance improve if we ignore sentences that describe the *context* in which the searcher watched the movie? Or, does performance improve if we ignore sentences that convey uncertainty?

Our results provide insights about the behaviors of searchers in TOT states and the effectiveness of existing IR systems in responding to TOT information requests. Our results suggest that searchers in TOT states tend to leverage memories about the movie itself (e.g., scenes, characters, locations) and the contexts in which they viewed the movie (e.g., time, place, and even concurrent external world events). The presence of contextual information in TOT requests is consistent with previous results from the personal information management (PIM) literature [10–12, 17]. While searchers may not recall the exact keywords in an email, they may recall things that were happening when the email was received. Additionally, our results suggest that searchers use a variety of tactics that are not supported by conventional search systems, including multi-hop reasoning (e.g., comparisons to other movies), self-reflective descriptions of previous search attempts, and expressions of uncertainty. In terms of retrieval performance, our results suggest that current IR systems can successfully leverage descriptions of the movie, but not descriptions of the context. Furthermore, they are surprisingly robust to expressions of uncertainty. We discuss opportunities for future research to develop algorithmic solutions to resolve these types of information needs.

2 RELATED WORK

Our research builds on three areas of prior work. First, our work builds on psychology research on long-term memory, asking what people remember and how they convey memories. Similarly, our work builds on prior information retrieval research aimed to support information *re-finding* in situations where a searcher has lapses in memory. Much of this research has been done in the context of *personal information management* (PIM). Finally, prior IR research has also studied how information requests aimed at human intermediaries differ from requests aimed at a search system.

Long-term Memory. Long-term memory plays an important role in information re-finding. It is generally accepted that human memory is transient. Research in psychology has studied how different factors contribute to lapses in long-term memory. For example, research clearly shows that long-term memory degrades with time (i.e., decay theory [35]) and by an individual engaging with new tasks and information objects (i.e., interference theory [2]). Additionally, memory degrades quicker when an individual did not explicitly aim to remember the item in question (i.e., poor encoding [12]). In the context of movies, recall may degrade with time and by a searcher engaging with other (perhaps similar) movies. Additionally, gaps in memory may occur simply because the searcher did not aim to remember the movie in the long term.

Psychology research has studied, not only *how much* is remembered, but also *what* is remembered. For example, research shows that people tend to forget precise details and remember high-level characteristics or "gists" [34]. In the context of movies, a searcher may forget the name of a character, but remember their personality. Additionally, models of long-term memory distinguish between *declarative* memory (i.e., remembering bits of information) versus *procedural* memory (i.e., remembering skills). Furthermore, declarative memory is sub-divided into *semantic* memory (i.e., memory about inherent characteristics of an information item) versus *episodic memory* (i.e., memory about previous engagements with the item) [36, 39]. Episodic memory can be viewed as "autobiographical" and deals with subjective experiences.

[2] https://www.labnol.org/internet/find-name-of-songs/12316/
[3] https://irememberthismovie.com/

Previous studies about *how* people recall have demonstrated consistent strategies used by individuals. For example, when presented with the photograph of a famous person, individuals leverage information about the person (e.g. profession, places) in order to recall their name [41]. This is an example of the 'tip of the tongue' phenomenon, 'a state in which one cannot quite recall a familiar word but can recall words of similar form and meaning' [6, 37]. Compared to other information seeking tasks, both the information retrieval [12] and cognitive psychology [29] literature indicate that searchers in TOT states exhibit more frustration at not knowing the answer and more satisfaction when the answer is revealed.

Interestingly, psychology research has used movies to study long-term memory in a controlled laboratory setting. As noted in Furman et al. [14, p457], using movies to study long-term memory is appealing because they can simulate "aspects of real-life experiences by fusing multimodal perception with emotional and cognitive overtones." Furman et al. [14] conducted a study in which participants watched a 30-minute movie and completed tests to measure long-term memory at different times, ranging from 3 hours to 9 months after watching the movie. As one might expect, test performance *and* self-reported confidence in the test answers degraded over time. Interestingly, however, performance degraded differently for test questions about different aspects of the movie. For example, performance degraded quicker for questions that asked about specific details (e.g., verbatim quotes) than questions that asked about themes and scenes involving social interactions.

Memory and Personal Information Management (PIM). PIM research studies how people manage and access their personal information (e.g., files, emails, photos, etc.). Memory plays an important role in PIM. Much PIM research has studied the importance of episodic memory (versus only semantic memory) during search and re-finding. In other words, PIM studies suggest that systems should support searching and re-finding using contextual cues. For example, a searcher may not remember the contents of an email in order to create an effective keyword query but may remember the day the email was received or events that happened that day.

Dumais et al. [10] evaluated the "Stuff I've Seen" system and found that "last modified date" was the most widely used contextual cue for filtering search results. Elsweiler et al. [12] conducted a diary study focused on participants' everyday memory problems and strategies used to overcome these. Results confirmed the importance of episodic memory to support information re-finding—participants often forgot details about the item itself, but remembered contextual details about when the item was last used (e.g., the task they were trying to accomplish when they last engaged with an item). Elsweiler et al. [11] evaluated a search system for managing, tagging, and accessing personal photographs. The system allowed searchers to search (and re-find) based on episodic memories. Results found that participants often searched based on *multiple* contextual cues (e.g., time, place, event-type, etc.). Hwang et al. [17] evaluated a system to tag bookmarked pages with contextual information. For example, participants could tag bookmarks based on time, location, and the task the participant was working on when the bookmark was made. Results found that contextual cues were effective in helping participants re-find bookmarks. Additionally, the task associated with a bookmark was the most recalled contextual cue.

In the context of life-logging, prior work also suggests that episodic memories (combined with semantic memories) can help searchers find information within their own *human digital memory* (HDM) repositories [13]. Similarly, Kelly et al. [22] compared the effectiveness of queries containing only content information (i.e., leveraging semantic memory) versus queries containing both content and context information (i.e., leveraging semantic and episodic memory). Queries combining content-related and contextual cues outperformed content-only queries.

Information Requests aimed to Humans. Wilson [40] proposed that information needs gradually evolve over four stages. First, a *visceral* need is one that cannot be expressed in words—there is a vague sense of unease that cannot be explained. Second, a *conscious* need is one that remains ambiguous (i.e., the searcher does not know what information is needed) but could be potentially resolved by talking to others. Third, a *formalized* need is one that can be communicated to others, but perhaps not to a search system. Finally, a *compromised* need is one that can be formulated using a specific interface or query language (e.g., by choosing specific keywords). In this paper, we study TOT information requests posed to human intermediaries. One might argue that these are cases where a searcher does not have the requisite knowledge to transition from a formalized state to an effective compromised state. To build better information retrieval systems, it is important to study not only *compromised* information needs (e.g., how people formulate keyword queries), but also *formalized* information needs (e.g., how people convey information needs to a human intermediary).

Arguello et al. [1] conducted a large-scale user study that compared information requests (for the same information needs) aimed to a search system versus a human intermediary. Additionally, they considered information needs associated with specific types of *extra-topical* relevance criteria (e.g., temporal-, geographical-, complexity-related criteria). Through qualitative analysis, the authors compared the search strategies adopted by searchers when conveying extra-topical relevance criteria during requests aimed at a human intermediary versus a search system. In the human intermediary condition, participants reported less difficulty producing their requests but adopted search strategies that deteriorated retrieval performance when the request was issued to a web search engine. For example, in the human intermediary condition, participants were more likely to convey what they *did not* want.

Kato et al. [20, 21] also studied how people formulate requests when the information need has a specific extra-topical dimension (e.g., domain knowledge). In this case, requests were aimed to a search system. Results found that participants often ignored the extra-topical dimension in their queries or used "indirect" strategies that work well with current search systems (e.g., using the query-term 'Wikipedia' to get results for a domain novice). Queries performed poorly when they explicitly mentioned the extra-topical dimension (e.g., 'simple explanation...').

More closely related to our research, Hagen et al. [15] curated a corpus of known-item questions posed to the Yahoo! Answers Q&A site. Most questions aimed at (re-)finding a website. However, some questions aimed at (re-)finding a previously experienced item (e.g., movie, book, song, band/musician, etc.). Interestingly, a qualitative analysis found that 240 questions (out of 2,755) contained

so-called "false memories", in which the asker provided incorrect information (e.g., mentioning the wrong actor when trying to recollect a movie). In a recent paper, Jørgensen and Bogers [18] report on qualitative analysis of TOT requests posed to the 'Tip of my Joystick' Reddit community, aimed at helping people re-find previously played video games. The coding scheme developed focused on similar phenomena as ours—visual characteristics (e.g., characters), audio characteristics (e.g., soundtrack), metadata characteristics (e.g., release date), comparisons with other video games, and characteristics of the *context* in which the asker previously engaged with the video game. The authors discuss controlled vocabularies necessary to support TOT requests for video games.

3 MOVIE IDENTIFICATION

Query-based search is an important component of movie streaming services such as Netflix and YouTube. However, existing systems tend to rely on metadata (e.g., title, actor, director, genre, time period) as opposed to plot or scene descriptions [16, 24, 27]. In our case, we focus on TOT information requests in which the searcher does not remember metadata information or the intent is not easily expressible as a keyword query.

Perhaps due to the failure of current search systems to support searchers in TOT states, several community question answering sites have gained popularity on the web. These sites allow searchers to pose a question, composed of two parts: a title for the question and a longer, free text description of the item. In response to a question, a conversational thread consists of community members asking clarifying questions and suggesting answers while searchers respond to questions and indicate when an answer is correct.

We collected questions posted to 'I Remember This Movie', a community question-answering site specifically designed for individuals seeking to identify movies that they have seen but whose title they cannot recall. Because we could not confirm that all questions in fact referred to real movies, we restricted our collection to questions with correct answers, as indicated by the searcher. This resulted in a set of 2,072 questions posted between 2013 and 2018. Of these, 762 correct answers included links to IMDb pages, which we recorded for further analysis.

4 ANALYSING TOT REQUESTS

In order to better understand TOT requests and address **RQ1**, we performed a qualitative analysis of the requests from the dataset described in Section 3. Our analysis operated at the granularity of sentences, to allow easier annotation and interpretability of results. We developed a qualitative coding scheme to be applied at the sentence level and then analysed these annotations on our corpus.

4.1 Coding Scheme

We segmented TOT requests into sentences using the Stanford NLP toolkit[4] and applied qualitative codes to all sentences from a randomly sampled set of 1,000 requests.

Our coding scheme was developed in a top-down and bottom-up fashion. First, after looking at the data, four of the authors agreed on *seven* broad categories for codes: *movie*, *context*, *previous search*, *social*, *uncertainty*, *opinion/emotion*, and *relative comparison*. Four of these categories (*movie*, *context*, *previous search*, and *social*) focus on the topic of the sentence. The sentence may describe the movie, the context in which the movie was seen, a previous attempt to find the movie, or merely be a social nicety. The remaining three categories (*uncertainty*, *opinion/emotion*, and *relative comparison*) relate to interesting phenomena that we noticed and wanted to capture in our coding scheme. We noticed sentences that express uncertainty about the information being conveyed, express opinions about an aspect of the movie, describe emotional states, and provide descriptions in relative terms by drawing comparisons.

After identifying these seven broad categories, codes were developed in a bottom-up fashion by two of the authors. Codes were developed in three phases. During the first phase, two of the authors independently developed their own codes and then met to discuss their codes and definitions. At this point, both authors developed a preliminary coding scheme (i.e., codes and definitions) using the union of their individual coding schemes. During the second phase, both authors independently coded 50 randomly sampled sentences. After measuring inter-annotator agreement, codes with low agreement (i.e., Cohen's $\kappa \leq .20$) were redefined, dropped, or combined with other codes. Finally, during the third phase, both of the authors independently coded 758 randomly sampled sentences. At this point, inter-annotator agreement for all codes was satisfactory. Our final coding scheme (described below) resulted in 34 codes. Of these, 2 codes (6%) had a Cohen's κ agreement at the level of 'fair', 5 (15%) at the level of 'moderate', 16 (47%) at the level of 'substantial', and 9 (26%) at the level of 'almost perfect' [25]. All codes were designed to not be mutually exclusive, meaning that sentences could be assigned zero, one, or more codes.

After developing and testing the reliability of our coding scheme, both coders coded 1,000 TOT information requests (500 each) comprised of 8,030 total sentences. Our coding scheme was comprehensive in terms of capturing a wide range of phenomena. As it turns out, all sentences coded were assigned at least one of our 34 codes.

4.2 Results

Our codes are described in Table 1 (content codes), Table 2 (context codes), and Table 3 (other codes). In each table, the first column contains the code label, the second column provides a definition, the third column illustrates an example sentence that was assigned the code, the fourth column shows the relative frequency of the code (i.e., the percentage of sentences associated with the code), and the fifth column shows the Cohen's κ agreement for the code. As previously mentioned, to test the reliability of our coding scheme, two of the authors independently coded the same set of 758 randomly sampled sentences. To measure inter-coder agreement, we computed the Cohen's κ agreement for each code independently. Cohen's κ ranges from -1 to +1 and measures agreement after correcting for agreement due to random chance. A value of -1 signals perfect disagreement, +1 signals perfect agreement, and 0 signals agreement no better than random. The relative frequencies (fourth column) in Tables 1–3 do not sum to 100% because codes were not mutually exclusive.

Movie. As expected, many sentences described characteristics of the movie itself. As shown in Table 1, these sentences include descriptions of the characters in the movie, a scene, a physical

[4] https://nlp.stanford.edu/software/

object, the movie's category/genre/tone, and the overall plot. This high-level dimension focuses on *semantic* memories—recollections about the movie itself.

Context. We also noticed that some sentences describe the context in which the searcher previously engaged with the movie. As shown in Table 2, these included references to when and where the searcher watched the movie, the medium (e.g., TV, movie theatre, etc.), who they watched it with, and even world events that were happening around the time period they watched the movie. This high-level dimension focuses on *episodic* memories—recollections about the context in which the searcher previously engaged with the movie. As noted in Section 2, episodic memories play an important role during information re-finding in the context of personal information management [10–12, 17]. For example, Dumais et al. [10] found that one might not remember the content of an email, but may remember when they received it or what was happening when they received it. The annotations based on our context codes suggest that episodic memories also play a role in TOT information needs.

The remaining codes are described in Table 3.

Previous Search. The previous search dimension focuses on references to failed attempts to identify the movie title. These include references to previous information sources consulted (e.g., search engines, websites, people), as well as descriptions of search strategies that were unsuccessful (e.g., searching through an artist's filmography).

Social. This category relates to sentences containing social niceties (e.g., please and thank you) [8]. Sentences were labeled at a high rate (10.77%) compared to other groups of annotations.

Uncertainty. Interestingly, we also noticed that searchers often used linguistic markers of uncertainty (e.g., "I vaguely remember that..."). The high rate of this annotation (35.37%) suggests that searchers in TOT states are self-aware when providing incomplete or unreliable information due to long-term memory degradation.

Opinion/Emotion. The opinion/emotion dimension focuses on the presence/absence of opinionated statements and references to emotional states. Here, we refer to opinions as judgements, critiques, or evaluations about some aspect of the movie. Conversely, we refer to emotional expressions as references to emotional states experienced by the searcher while watching the movie.

Relative Comparison. This dimension relates to whether the sentence conveys information in relative versus absolute terms. An absolute statement is one that conveys information without drawing comparisons that require background knowledge or additional informational in order to extract its full meaning. For example, "The man character is a blond, handsome man." is an absolute statement. Conversely, a relative statement is one that draws one or more comparisons that require additional information in order to extract the meaning of the statement. For example, "The main character looks like Brad Pitt." is a relative statement. Extracting the meaning of this sentence requires resolving Brad Pitt's physical features. We view relative comparisons as statements that require some degree of inference using background knowledge. In other words, relative comparisons require multi-hop reasoning.

4.3 Discussion

Our qualitative analysis of TOT requests reveal several important trends. These trends provide insights about: (1) the things people remember and (2) the things people decide to convey when attempting to resolve their TOT information needs.

What people remember. First and foremost, our analysis reveals that people remember characteristics of the movie (e.g., a scene, character, object) as well as characteristics of the context in which the movie was seen (e.g., time, place, physical medium, external events). In other words, it appears that searchers rely on both semantic *and* episodic memories when attempting to resolve a TOT information need.

Second, based on our coding scheme, searchers conveyed visual memories more than auditory memories. Our most frequent codes (>18%) involved visual memories (e.g., character, scene, object, location type). Codes associated with auditory memories (e.g., quotes, compares music, specific music) were much less frequent (< 2%). We see at least two possible explanations for this trend. One possibility is that visual memories are easier to communicate than auditory memories. In other words, perhaps searchers had plenty of auditory memories, but they decided to omit them in their TOT queries. Alternatively, it is possible that searchers had more visual memories than auditory ones. This explanation would be consistent with prior research that has found that visual memory is more robust (i.e., long-lasting) than auditory memory [3, 7].

Third, most of our frequent codes are related to things that exist in the physical world and can be perceived by the senses (e.g., character, scene, object, location type). Only one of our codes (i.e., tone) relates to an abstract characteristic of movies (e.g., dark, scare, fantasy). This trend also seems consistent with prior work that has found that memories of concrete characteristics are more robust than abstract ones [19].

What people say. In terms of what searchers communicate, our results suggest four important trends. First, searchers convey information that may be useful for a human intermediary (with domain and world knowledge), but potentially problematic for existing search systems that rely (partly or entirely) on keyword matching. In particular, about 10% of sentences contained descriptions of the context in which the searcher previously watched the movie. These references (based on episodic memories) may be helpful for an intelligent human searcher, but require some degree of inference using real-world knowledge. Table 2 provides some examples. For instance, mentioning that the movie was seen in a "film class" (*physical location*) implies that the movie is probably artistic or noteworthy; mentioning that "I was so young my parents made me turn it off" (*contextual witness*) implies that the movie is not appropriate for children; and mentioning that "I watched it alongside Hard Candy" (*concurrent events*) implies that the movie came out around 2005.

Secondly, it is interesting that searchers mentioned previous (failed) attempts to find the movie. To gain more insight about these references, we examined the codes with the highest degree of co-occurrence with our *previous search* code. Table 4 shows all codes with a positive point-wise mutual information (PMI) score with *previous search*. The first column provides the co-occurring

Table 1: *Movie Annotation.* Codes related to characteristics of the movie.

code	definition	examples	frequency	κ
Character	Describes a character.	The main protagonist is a 20-something girl with short hair, which is either blonde or brunette.	51.21%	0.766
Scene	Describes a scene.	Finally the real boyfriend appeared at the final scene in a cabin near some lake or sea and they try to kill each other.	36.53%	0.755
Object	Describes a tangible object in a scene.	They're in the car and they almost crash into this beast.	26.72%	0.750
Category	Describes the movie category (e.g., movie, tv movie, miniseries, etc.).	Live-action possibly made-for-TV.	25.07%	0.536
Location type	Describes a scene's location type.	The movie starts out with an American family who are staying in some Eastern European Castle with their young son.	18.36%	0.698
Plot summary	Describes the overall plot or premise.	This movie is about a young girl who marries early and has a baby boy.	10.82%	0.637
Release date	Describes timeframe of movie release.	I remember this horror movie from late 70s early 80s.	5.43%	0.854
Genre/tone	Describes genre or tone.	I think it was a romantic comedy of sorts.	5.39%	0.782
Visual style	Describes visual style (e.g., black and white, colour, CGI animation, etc.).	It was in English not subtitled and in colour.	4.73%	0.821
Language	Describes the language spoken.	The dialog in the movie was in Spanish.	2.89%	0.955
Regional Origin	Describes movie's region of origin.	I think it was an European movie and not in English.	1.72%	0.933
Specific location	Describes a scene's specific location.	I believe they were traveling to Louisiana to pick up a friend's body for a funeral.	1.58%	0.598
Quote/dialogue	Describes a quote from the movie.	The wife yells something along the lines of either 'look what you did' or 'look what you did to my husband'.	1.54%	0.766
Real person	Describes real person associated with movie (directly or indirectly).	The woman looked like Annie Clark .	1.21%	0.864
Camera angle	Describes camera action.	The jumping between scenes was also very strange.	0.95%	0.663
Singular timeframe	Describes timeframe.	I think it was made in either the 70s or 80s but the movie is set in the 20s or 30s.	0.71%	0.832
Multiple timeframe	Describes the passage of time in the movie.	Decades later the house that is above the tunnel is believed to be haunted.	0.63%	0.799
Fictional person	Describes fictional person associated with movie (directly or indirectly).	It was a scene with two adventures in a scene like Indiana Jones trapped captured by enemy forces.	0.62%	0.712
Actor nationality	Describes nationality or ethnicity associated with actor/actress.	She is a regular height woman also Caucasian slim and with red hair.	0.54%	0.499
Target audience	Describes movie's target audience.	Gadget packed action movie for Kids?	0.49%	1.000
Compares music	Describes movie's soundtrack.	I remember there was lots of nice electronic music but what was the title of the movie?	0.32%	0.888
Specific music	Describes specific song in the movie.	The mother makes her living from singing in small joints at some point she sings a version of "Looking for the Heart of Saturday Night".	0.15%	0.666

Table 2: *Context Annotation.* Codes related to characteristics of the context surrounding the searcher's previous engagement with the movie.

code	definition	examples	frequency	κ
Temporal context	Describes when the movie was seen, either in absolute terms (e.g., around 2008) or relative terms (e.g., when I was a kid).	I rented this film in the early 2000's.	8.58%	0.783
Physical medium	References the physical medium associated with watching the movie (e.g., TV, theatre, VHS, etc.)	I remember it was like in the 2000's and it was on the tele.	5.42%	0.855
Cross media	Describes exposure to movie through different media (e.g., trailer, DVD cover, poster, etc.)	One of its posters shows a man waving a sheet of white cloth.	1.06%	0.542
Contextual witness	Describes other people involved in the movie-watching experience.	I remember it was on television and I was so young my parents made me turn it off.	0.76%	0.621
Physical location	Describes physical location where movie was watch.	I watched this movie in my film class a couple years ago.	0.72%	0.621
Concurrent events	Describes events relevant to time period when movie was watched.	I've seen it around 2006 (I know cause I watched it alongside Hard Candy).	0.14%	1.000

Table 3: *Other Annotations.* Codes associated with previous search attempts, social niceties, uncertainty, opinions, emotions, and relative comparisons.

code	definition	examples	frequency	κ
Previous search	Describes a previous attempt to find the movie title.	I tried to find it using google, searched a number of databases with sci-fi movies from 1960s-1990s with no success.	1.48%	0.811
Social	Communicates a social nicety	If you could help at all I'd really appreciate it!	10.77%	0.735
Uncertainty	Conveys uncertainty about information described.	It was a foreign film I think either French or German, but I could be wrong.	35.37%	0.512
Opinion	Conveys an opinion or judgement about some aspect of the movie.	Its pretty confusing all the way to the end when there's only one surviving woman and then she is sat in the same room with this monster.	2.09%	0.341
Emotion	Conveys how the movie made the viewer feel.	It was the first movie that kept me awake at night.	0.46%	0.283
Relative comparison	Describes a characteristic of the movie in relative (vs. absolute) terms.	One of the detectives is young laid back kinda like Kevin Bacon or Gary Sinise but looking through their filmography I could not find the movie.	3.01%	0.701

code, the second column provides an explanation, the third column provides the PMI score, and the fourth column provides an example sentence. Four codes had positive PMI values: *real person*, *relative comparison*, *genre/tone*, *release date*. Based on the examples, searchers contributed potentially useful *negative* evidence that may help someone identify the movie. The examples point to failed attempts to find the movie by searching for all movies from a certain actor/actress, all episodes of a specific series, and all movies from a given genre and release date. This information might help someone define the search space (i.e., rule out particular alternatives). This trend suggests that systems to support TOT information needs may benefit from accommodating negative feedback.

Third, relative comparisons (versus absolute statements) were found in 3% of all sentences. This result suggests that it is often easier for someone to draw a comparison (e.g., "looks like Kevin Bacon") than to describe someone or something in absolute terms. Table 5 shows all codes with a positive PMI score with respect to *relative comparison*. Interestingly, common relative comparisons included comparisons with fictional characters (e.g., "tarzan"), real people (e.g., Kirsten Dunst), time periods (e.g., Victorian-esque), genres/tones (e.g., "let's hunt humans for fun type movie"), regional origins (e.g., "American-style"), and specific locations (e.g., "Grand Canyon-esque"). This trend suggests that systems to support TOT information needs may need to accommodate comparisons between people and other movie attributes. Prior work has found that information retrieval systems perform poorly on queries containing relative (versus absolute) statements [1].

Finally, it is noteworthy that expressions of uncertainty were so common. Roughly 35% of all sentences contained expressions of uncertainty. In linguistics, hedging allows speakers and writers to signal caution or probability versus full certainty. Again, Table 6 shows all codes with a positive PMI score with respect to *uncertainty*. Searchers expressed uncertainty about the movie's release date, the regional origin, an actor/actress in the movie, the lyrics of a song, the movie's timeframe, the target audience, a specific location, a specific musical piece, and even the temporal context when the movie was watched. This trend suggest that systems to support TOT information may need to deal with (un-)certainty. While information retrieval systems have not been designed to model uncertainty of a searcher's input, hedging has been incorporated in other types of systems (e.g., speech-based tutoring systems [31]).

5 ANALYSING AUTOMATIC RETRIEVAL FOR TOT REQUESTS

In addressing **RQ1** in the previous section, we found that searchers in TOT states used a variety of strategies to express their information need. In this section, we will be focusing on **RQ2** and how conventional automatic retrieval systems perform in response to these requests. Specifically, we are interested in the ability of retrieval systems to support the various tactics adopted by searchers in TOT states. To this end, we conduct per-code ablation experiments using a standard retrieval system (Okapi BM25 [32]).

5.1 Methods

In order to accommodate identification experiments, we need: (1) a collection of *search requests* and the relevant item for each request, (2) a *corpus* where each document is associated with a unique item, (3) a *retrieval system*, and (4) an appropriate *evaluation metric*. We describe the search requests for our retrieval experiments in Section 3. Because this is an identification task and we only selected TOT requests with a correct answer, each request had exactly one answer (i.e., the correct movie title).

For the corpus, we wanted to assemble a set of indexable items that would be amenable to free text retrieval and the movie identification task. Long form movie plot descriptions provide a text-rich representation of movies. We extracted the plot description for each movie in Wikipedia. We used the WikiPlot[5] code to extract plots from a 2019 dump of Wikipedia. We indexed the WikiPlot collection with the Indri retrieval system [38], removing stopwords using the Indri stopword list and stemming using the Krovetz algorithm [23]. Each plot description was an average length of 328.9 words after removing stopwords. Consistent with indexing, the query terms are also stemmed during retrieval.

In order to associate each request with a relevant document in the Wikipedia corpus, we restricted both the requests and the documents to those associated with a unique IMDb identifier. Several of the correct answers in the 'I Remember this Movie' dataset included links to the IMDb page. Additionally, many Wikipedia movie pages include a reference to the movie's IMDb page. Filtering for those Wikipedia entries resulted in 69,132 documents and 339 requests with matches in that corpus. Each request had exactly one relevant document in the collection.

As a retrieval method, we adopted Okapi BM25 [32], allowing a reproducible standard retrieval algorithm. We used a 20% randomly sampled subset of the training set to tune the BM25 parameters.

Finally, we adopted "success at ten" as our evaluation metric. For a given query, this metric is defined as 1 if the correct movie was returned in the top ten positions; 0 otherwise. This metric reflects the searcher's recall orientation more than mean reciprocal rank, a metric often used for question answering.

To study the relative usefulness of different phenomena (i.e., qualitative codes) present in TOT requests, we conducted ablation experiments. We adopted the following protocol for this study. For each code, we first computed the retrieval performance using all the requests that included at least one sentence with that code. Then, for each of these requests, we computed the retrieval performance after removing all sentences with the code. If performance degrades after removing all sentences with the code, it means that those sentences contributed information that the retrieval algorithm was able to harness to improve results (on average). Conversely, if performance improves, it means that those sentences included content that degraded retrieval performance. Because some of our codes occur infrequently, we focus only on codes that occurred in $> 20\%$ of all TOT requests.

5.2 Results

Averaging across *all* 339 requests, we found that only 13.27% of requests placed the relevant document above the tenth position and 55.16% retrieved the relevant document above the 1000th position. We present the results for our ablation experiments in Table 7.

[5] https://github.com/markriedl/WikiPlots

Table 4: Codes with highest pointwise mutual information (PMI) with 'Previous search'.

Co-occurring Code	Explanation	PMI	Example
Real person	Searched by potential actor/actress.	3.564	For some reason I remember it as Julia Stiles but I looked at IMDb and nothing on her filmology page rings any bells.
Relative comparison	Searched by comparing with similar/related items.	2.617	My grandparents were fairly "proper" people so I expected this to be an episode of Masterpiece Mystery or Poirot but I can't find it.
Genre/tone	Searched by genre/tone.	0.64	I've gone through countless lists like "50 weird SciFi movies from the 80's" and still nothing.
Release date	Searched by release date.	0.308	I tried manually browsing wikipedia page of scifi movies from 70ies up to now but I can't seem to find it.

Table 5: Codes with highest pointwise mutual information (PMI) with 'Relative comparison'.

Co-occurring Code	Explanation	PMI	Example
Fictional person	Comparisons with a fictional character.	4.408	He looked similar to Tarzan, but he wore pants and had some kind of weapon strapped across his back.
Real person	Comparisons with other actors/actresses or comparisons with the artistic styles of other writers/directors.	3.976	For some reason, I swear Kirsten Dunst was in this movie and keep thinking it has something to do with The Virgin Suicides but it is not that movie and I can not find it when searching Kirsten Dunst.
Previous search	Comparisons with other movies or artists in the context of a prior search attempt.	2.617	This'll be an easy one for you guys I'm sure, but Googling just brings up "The Craft" and "Slugs".
Opinion	Opinionated comparisons.	2.614	And it has that 70's horrible sound quality, especially when someones screaming or he's doing his creepy laugh.
Singular timeframe	Comparisons with temporal periods.	2.027	It was set in a Victorian-esque setting with horse drawn carriages.
Genre/tone	Comparisons with styles of other films.	1.720	It was definitely a "Let's hunt humans for fun" type movie.
Cross media	Comparisons with media associated with other movies.	1.450	The cover seemed almost like a National Lampoon cover.
Regional origin	Comparisons with movies from specific origin.	1.114	Also, it was translated to Turkish but the movie itself looked very American-style children's movie.
Specific location	Comparisons between a scene location and a real one.	1.064	The river looked "Grand Canyon-esque" we find out one of the girls mom gets beat up really bad by her husband and at the end he gets arrested.

Table 6: Codes with highest pointwise mutual information (PMI) with 'Uncertainty'.

Co-occurring Code	Explanation	PMI	Example
Release date	Uncertainty about release date.	1.153	I remember a movie about a "super bus", I think it was in the late 70's.
Regional origin	Uncertainty about the movie's regional origin.	0.930	Can't remember the country of origin but I believe it was Scandinavian.
Real person	Uncertaintly about a person associated with the movie.	0.899	The girl had a "smiling dimple" that reminded me of Sarah Michelle Gellar but I'm absolutely unsure.
Compares music	Uncertainty about movies sountrack.	0.886	The lyrics were something like "it's a crazy world" and in the video there are lots of people doing crazy stuff and there is a guy kicking a (fake) dog over a balcony.
Singular timeframe	Uncertainty about the movie's timeframe.	0.836	I believe it was based in New York and during the 80's.
Target audience	Uncertaintly about the target audience.	0.799	I believe this is a kids movie I saw on TV in the mid to late nineties.
Specific location	Uncertainty about locations in the movie.	0.758	All I remember (correctly I hope) is that it was a movie maybe 70's comprised of a number of unrelated scenes all set in Europe.
Specific music	Uncertainty about a specific song in the movie.	0.721	It was like that song with the rabbit and the three blind mice .
Temporal context	Uncertainty about the time when movie was watched.	0.693	Here is one I saw long ago >>20 years on TV probably a "Creature Double Feature".

In general, sentences descriptive of the content of the movie, when removed, resulted in 9.44% of requests no longer retrieving the relevant movie in the top ten positions. However, this impact was not uniform across all content categories. The most substantial influence came from descriptions of characters, objects, scenes, locations, and plot summaries; the remaining descriptive types tended to have minimal impact on performance, roughly resulting in negligible changes in success rate. Sentences with context information (e.g., when and where a movie was seen) similarly had negligible impact on performance. Finally, sentences expressing uncertainty *helped* performance. In other words, we observed an increase in failure of requests after removing sentences where searchers expressed uncertainty. Neither sentences containing relative comparisons nor those containing social niceties substantially affected retrieval performance.

5.3 Discussion

Our experiments suggest that there is substantial room for improving systems to support TOT requests. We observed this despite the alignment between our corpus of plot descriptions and the dominant searcher strategy of describing the movie content. As such, we believe that richer, more granular representations of salient or memorable content should improve retrieval performance. We note that what is memorable may in fact be different from a straightforward plot description.

Importantly, we observed systematic variation in retrieval performance across different strategies employed by searchers in TOT states.

Table 7: Ablation experiments for each annotation label. 'Frequency' denotes the percentage of TOT requests containing sentences associated with the code. 'All' denotes performance by using the entire TOT request; 'ablated' denotes performance by omitting sentences associated with the code. Columns 'absolute' and 'relative' denote the difference in performance in absolute terms and percent increase/decrease, respectively. Rows sorted within category by relative difference with respect to the baseline. Larger drops in performance indicate important sentence types.

		success@10		difference	
	frequency	all	ablated	absolute	relative
Movie (all)	100%	0.1327	0.0383	-0.0944	-71.1%
Character	98.2%	0.1351	0.036	-0.0991	-73.4%
Object	82.3%	0.1434	0.0466	-0.0968	-67.5%
Scene	89.7%	0.1382	0.0625	-0.0757	-54.8%
Location type	73.7%	0.148	0.096	-0.052	-35.1%
Plot summary	61.7%	0.1531	0.1148	-0.0383	-25.0%
Category	92.0%	0.1346	0.1314	-0.0032	-2.4%
Genre/tone	34.8%	0.0763	0.0763	0	0.0%
Release date	43.4%	0.0952	0.102	0.0068	7.1%
Visual style	34.2%	0.1552	0.1897	0.0345	22.2%
Language	23.9%	0.1111	0.1358	0.0247	22.2%
Context (all)	64.9%	0.1227	0.1409	0.0182	14.8%
Temporal context	57.8%	0.1327	0.148	0.0153	11.5%
Physical medium	35.7%	0.124	0.1488	0.0248	20.0%
Other					
Uncertainty	88.5%	0.1333	0.1067	-0.0266	-20.0%
Relative comparison	20.6%	0.0571	0.0571	0	0.0%
Social	51.6%	0.0686	0.0743	0.0057	8.3%

Because we indexed detailed movie plots, a searcher's description of characters, objects, scenes, and locations allowed for effective retrieval. However, our results also suggest the importance of content metadata in supporting TOT requests. Request sentences referring to genre/tone, category, release date, language, and visual style all refer to information about the movie, rather than the content of the movie. While the lack of effectiveness of types of strategies is likely attributable to missing metadata, we also note that, more generally, metadata descriptions may be too coarse (i.e., applicable to more than one, if not many, movies to result in effectiveness improvements over more precise information specified in a plot description).

Contrary to our expectations, our results demonstrate that sentences expressing uncertainty did *not* degrade performance. This indicates that even when the searcher may be unsure, the stated movie description may be accurate and useful for retrieval. For example, for the sentence "I saw the movie somewhere around 1986-88 on TV about young (high school) musician boy who played electronic music on a keyboard", we find term matches with the correct movie plot description that also contains the phrase "electronic music", despite uncertainty about the year. Similarly, sentences that make relative comparisons can also often include important keywords that match with the target movie plot description. For example, the sentence "I don't know if this Laika was the original Laika who travelled to the space in the 50s or the name is in honor of the real astronaut dog" contains the terms "space" and "dog" that matches with the target movie plot description. The inclusion of this sentence in the query formulation on the whole seems to have a positive effect on retrieval—in spite of terms like "Laika" appearing in eleven other plot descriptions in our collection.

We were concerned that algorithms may be brittle in the presence of unsupported strategies like metadata or contextual information. Fortunately, in both cases, we found minimal deterioration in performance after removing sentences with these codes.

Our retrieval experiments were conducted in the context of standard Okapi BM25 model. In future work, it may be interesting to revisit similar research questions in the context of other more sophisticated IR models, such as those that learn latent representations of text [30] or operate over structured content [33, 42].

6 CONCLUSION AND FUTURE WORK

In this paper, we describe tip of the tongue known-item retrieval, a class of item identification tasks where the searcher has previously experienced or consumed the item but cannot recall an identifier. Previous research demonstrates that TOT states can be especially frustrating to searchers and, as a result, have led to the creation of community question-answering sites around these needs, covering cultural objects such as movies, music, and books. Our qualitative coding of a set of TOT requests indicate that searchers employ a variety of information-seeking strategies, including semantic and episodic memories of previous experiences with the item. Moreover, searchers leverage more sophisticated constructs such as multi-hop reasoning, self-reflective descriptions of previous search attempts, and expressions of uncertainty. Inspite of the sophistication of these techniques, we found that automatic retrieval was largely unaffected by the presence of these operations.

We believe that TOT requests reflect an important and open area of information retrieval research. We used movie identification as a case study and several of our observations may or may not exist in domains such as music, books, or other media. While specific codes we developed may need to be adapted, we suspect that there are more abstract, general behaviors repeated by searchers across other domains. At that, the range of tactics employed during TOT states—perhaps due to frustration—makes this a rich context within which to observe searcher behavior in controlled environments. This would require the adaptation of TOT elicitation techniques from the cognitive psychology domain to the information retrieval context [5].

From an algorithmic perspective, supporting document representations that are both comprehensive (i.e., including detailed descriptions and all metadata) and amenable to more elaborate search strategies will go a long way toward satisfying TOT needs. As noted in our results, this may require better understanding the distinction between descriptive representations and those biased toward memory-salience. At the same time, the integration of personal information management and life logging techniques will be necessary for responding to contextual information conveyed by searchers.

REFERENCES

[1] Jaime Arguello, Bogeum Choi, and Robert Capra. 2018. Factors Influencing Users' Information Requests: Medium, Target, and Extra-Topical Dimension. *ACM Transactions of Information Systems* 36, 4 (2018), 37.

[2] G. H. Bower, S. Thompson-Schill, and E. Tulving. 1994. Reducing Retroactive Interference: An Interference Analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 1, 20 (1994), 51–66.

[3] Timothy F. Brady, Talia Konkle, George A. Alvarez, and Aude Oliva. 2008. Visual Long-term Memory has a Massive Storage Capacity for Object Details. *Proceedings of the National Academy of Sciences* 105, 38 (2008), 14325–14329.

[4] Andrei Broder. 2002. A Taxonomy of Web Search. *SIGIR Forum* 36, 2 (2002), 3–10.

[5] A S Brown. 1991. A Review of the Tip-of-the-tongue Experience. *Psychol Bull* 109, 2 (Mar 1991), 204–223.

[6] Roger Brown and David McNeill. 1966. The "Tip of the Tongue" Phenomenon. *Journal of Verbal Learning and Verbal Behavior* 5, 4 (1966), 325 – 337.

[7] Michael A. Cohen, Todd S. Horowitz, and Jeremy M. Wolfe. 2009. Auditory Recognition Memory is Inferior to Visual Recognition Memory. *Proceedings of the National Academy of Sciences* 106, 14 (2009), 6008–6010.

[8] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL, 250–259.

[9] Doug Downey, Susan Dumais, Dan Liebling, and Eric Horvitz. 2008. Understanding the Relationship between Searchers' Queries and Information Goals. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM '08)*. ACM, 449–458.

[10] Susan Dumais, Edward Cutrell, JJ Cadiz, Gavin Jancke, Raman Sarin, and Daniel C. Robbins. 2003. Stuff I've Seen: A System for Personal Information Retrieval and Re-Use. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR '03)*. ACM, 72–79.

[11] David Elsweiler, Ian Ruthven, and Christopher Jones. 2005. Dealing with Fragmented Recollection of Context in Information Management. In *Proceedings of the 2005 Context-Based Information Retrieval Workshop in Conjunction with the Fifth International and Interdisciplinary Conference on Modeling and Using Context (CIR @ CONTEXT '05)*, Vol. 151.

[12] David Elsweiler, Ian Ruthven, and Christopher Jones. 2007. Towards Memory Supporting Personal Information Management Tools. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 924–946.

[13] Marguerite Fuller, Liadh Kelly, and Gareth Jones. 2008. Applying Contextual Memory Cues for Retrieval from Personal Information Archives. In *Proceedings of the 2008 Personal Information Management Workshop (PIM '08)*.

[14] Orit Furman, Nimrod Dorfman, Uri Hasson, Lila Davachi, and Yadin Dudai. 2007. They Saw a Movie: Long-term Memory for an Extended Audiovisual Narrative. *Learning & memory* 14, 6 (2007), 457–467.

[15] Matthias Hagen, Daniel Wägner, and Benno Stein. 2015. A Corpus of Realistic Known-Item Topics with Associated Web Pages in the ClueWeb09. In *Advances in Information Retrieval*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). Springer International Publishing, 513–525.

[16] Christine Hosey, Lara Vujović, Brian St. Thomas, Jean Garcia-Gathright, and Jennifer Thom. 2019. Just Give Me What I Want: How People Use and Evaluate Music Search. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, 1–12.

[17] Hyeon Kyeong Hwang, Eelco Herder, and Marco Ronchetti. 2017. A Link Back to MemoryLane: The Role of Context in Bookmarking and Refinding. In *Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter*. ACM, 8:1–8:10.

[18] Ida Kathrine Hammeleff Jørgensen and Toine Bogers. 2020. "Kinda like The Sims... But with Ghosts?": A Qualitative Analysis of Video Game Re-Finding Requests on Reddit. In *International Conference on the Foundations of Digital Games*. ACM.

[19] Marcel Just and Hiram Brownell. 1974. Retrieval of Concrete and Abstract Prose Descriptions from Memory. *Canadian Journal of Psychology/Revue canadienne de psychologie* 28 (1974), 339–350.

[20] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. 2014. Cognitive Search Intents Hidden behind Queries: A User Study on Query Formulations. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, 313–314.

[21] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, and Katsumi Tanaka. 2014. Investigating Users' Query Formulations for Cognitive Search Intents. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '14)*. ACM, 577–586.

[22] Liadh Kelly, Yi Chen, Marguerite Fuller, and Gareth J. F. Jones. 2008. A Study of Remembered Context for Information Access from Personal Digital Archives. In *Proceedings of the Second International Symposium on Information Interaction in Context*. ACM, 44–50.

[23] Robert Krovetz. 1993. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '93)*. ACM, 191–202.

[24] Sudarshan Lamkhede and Sudeep Das. 2019. Challenges in Search on Streaming Services: Netflix Case Study. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. ACM, 1371–1374.

[25] J. R. Landis and G. G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977), 159–174.

[26] Jin Ha Lee, Allen Renear, and Linda C. Smith. 2006. Known-Item Search: Variations on a Concept. *Proceedings of the American Society for Information Science and Technology* 43, 1 (2006), 1–17.

[27] Jingjing Liu, Scott Cyphers, Panupong Pasupat, Ian McGraw, and Jim Glass. 2012. A Conversational Movie Search System Based on Conditional Random Fields. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH '12)*.

[28] Qiaoling Liu, Eugene Agichtein, Gideon Dror, Yoelle Maarek, and Idan Szpektor. 2012. When Web Search Fails, Searchers become Askers: Understanding the Transition. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '12)*. 801–810.

[29] Janet Metcalfe, Bennett L. Schwartz, and Paul A. Bloom. 2017. The Tip-of-the-tongue State and Curiosity. *Cognitive Research: Principles and Implications* 2, 1 (2017), 31.

[30] Bhaskar Mitra and Nick Craswell. 2018. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval* 13, 1 (2018), 1–126.

[31] Heather Pon-Barry, Karl Schultz, Elizabeth Owen Bratt, Brady Clark, and Stanley Peters. 2006. Responding to Student Uncertainty in Spoken Tutorial Dialogue Systems. *International Journal of Artificial Intelligence in Education* 16, 2 (2006), 171–194.

[32] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[33] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple BM25 extension to multiple weighted fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM '04)*. 42–49.

[34] David C. Rubin. 1977. Very Long-term Memory for Prose and Verse. *Journal of Verbal Learning and Verbal Behavior* 16, 5 (1977), 611 – 621.

[35] David C. Rubin and A. E. Wenzel. 1996. One Hundred Years of Forgetting : A Quantitative Description of Retention. *Psychological Review* 3, 103 (1996), 734–760.

[36] Michael D. Rugg and Edward L. Wilding. 2000. Retrieval Processing and Episodic Memory. *Trends in Cognitive Sciences* 4 (2000), 108–115.

[37] Bennett L. Schwartz and Janet Metcalfe. 2011. Tip-of-the-tongue (TOT) States: Retrieval, Behavior, and Experience. *Memory & Cognition* 39, 5 (Jul 2011), 737–749.

[38] Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. 2005. Indri: A Language Model-based Search Engine for complex Queries. In *Proceedings of the International Conference on Intelligent Analysis*, Vol. 2. Citeseer, 2–6.

[39] Endel Tulving. 1993. What Is Episodic Memory? *Current Directions in Psychological Science* 2, 3 (1993), 67–70.

[40] Tom Wilson. 1999. Models in Information Behaviour Research. *Journal of Documentation* 55 (07 1999).

[41] A. Daniel Yarmey. 1973. I Recognize your Face but I can't Remember your Name: Further Evidence on the Tip-of-the-tongue Phenomenon. *Memory & Cognition* 1, 3 (1973), 287–290.

[42] Hamed Zamani, Bhaskar Mitra, Xia Song, Nick Craswell, and Saurabh Tiwary. 2018. Neural ranking models with multiple document fields. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (WSDM '18)*. 700–708.