Active tag recommendation for interactive entity search : Interaction effectiveness and retrieval performance

Ruotsalo, Tuukka

2022-03

Ruotsalo , T , Weber , S & Gajos , K Z 2022 , ' Active tag recommendation for interactive entity search : Interaction effectiveness and retrieval performance ' , Information Processing and Management , vol. 59 , no. 2 , 102856 . https://doi.org/10.1016/j.ipm.2021.102856

http://hdl.handle.net/10138/342308
https://doi.org/10.1016/j.ipm.2021.102856

Information Processing and Management 59 (2022) 102856

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Active tag recommendation for interactive entity search: Interaction effectiveness and retrieval performance

Tuukka Ruotsalo [a,b,*], Sean Weber [b], Krzysztof Z. Gajos [c]

[a] Department of Computer Science, University of Copenhagen, Denmark
[b] Department of Computer Science, University of Helsinki, Finland
[c] Harvard School of Engineering and Applied Sciences, MA, USA

ARTICLE INFO

Keywords:
Tag recommendation
Active learning
Information retrieval
Search user interfaces
User study

ABSTRACT

We introduce *active tag recommendation* for interactive entity search, an approach that actively learns to suggest tags from preceding user interactions with the recommended tags. The approach utilizes an online reinforcement learning model and observes user interactions on the recommended tags to reward or penalize the model. Active tag recommendation is implemented as part of a realistic search engine indexing a large collection of movie data. The approach is evaluated in task-based user experiments comparing a complete search system enhanced with active tag recommendation to a control system in which active tag recommendation is not available. In the experiment, participants (N = 45) performed search tasks on the movie domain and the corresponding search interactions, information selections, and entity rankings were logged and analyzed. The results show that active tag recommendation (1) improves the ranking of entities compared to written-query interaction, (2) increases the amount of interaction and effectiveness of interactions to rank entities that end up being selected in a task, and (3) reduces, but does not substitute, the need for written-query interaction (4) without compromising task execution time. The results imply that active learning for search support can help users to interact with entity search systems by reducing the need for writing queries and improve search outcomes without compromising the time used for searching.

1. Introduction

Many complex search tasks involve searching entities (Balog, 2018). While entity search has often been studied in an ad-hoc search setting (Balog & Neumayer, 2013; Balog et al., 2011), in the real-world search processes targeting to find entities are often interactive and characterized by system support that allows users to reformulate their queries and adjust their goals and information needs as new information is explored. To this end, researchers and commercial search engine providers are continuously introducing new approaches to enable functional search results or interactive widgets, aiming to assist users in exploring information without abandoning the search user interface. Interactive entity search, where entities are presented to the users as interactive and functional affordances to direct the search, can help searchers navigate diversified results and support exploratory search by highlighting relevant entities associated with a given user query (Bota et al., 2016; Ruotsalo et al., 2018).

On the social web, entities are often described by tags that users have associated with the entities and finding entities require users to use the tags in an interactive search process where users learn about the domain by searching and reflecting on the acquired information (Jones & Klinkner, 2008; Raman et al., 2014; Ruotsalo et al., 2014). This problem, however, is not limited to Web search

* Corresponding author at: Department of Computer Science, University of Copenhagen, Denmark.
E-mail address: tr@di.ku.dk (T. Ruotsalo).

https://doi.org/10.1016/j.ipm.2021.102856
Received 8 October 2021; Received in revised form 26 November 2021; Accepted 23 December 2021
Available online 10 February 2022

or social media, but it applies to any information access system that involves structured information describing entities. For example, a user looking for a movie to watch may be unfamiliar with the movies, genres, or topical structure available in the collection and would benefit from system guidance to filter and rank the information on different dimensions that are used in tagging the movie content. Moreover, the needs of the user may evolve as the search session progresses and information on related entities and tags become relevant in subsequent iterations as data are explored.

Interactions with conventional search engines mostly rely on typing queries or activating pre-defined filters or query suggestions. This promotes reactive user involvement as users only activate system suggestions to orient and direct their search as opposed to being actively involved in the query prediction process (Ruotsalo et al., 2018; Teevan et al., 2004). To this end, systems that assist the user with filtering recommendations, such as tags, dynamic facets, or query suggestions, often draw their predictions directly from information associated with the present query or the set of documents presently available for the user (Basu Roy et al., 2008; Dash et al., 2008; Dimitrov et al., 2018; Kammerer et al., 2009). That is, they are limited in the scope of the already retrieved information and allow limited exploration outside of the initial query scope. On the other hand, present approaches that do not rely on present query scope, such as more classic faceted search, require a faceted classification structure to be available.

A popular technique that supports interactive search and information exploration without the need for existing faceted classifications is tag recommendation (Kammerer et al., 2009). Tag recommendation harnesses the power of socially created tags that are associated with searchable entities. Tag recommendation has increasingly become a method for web users to organize and search online content. Many present web applications promote the use of tags to index and share content. Consequently, socially shared tags are used to search and find content, and already early research proposed social tagging as a good basis for enhancing entity search capabilities (Kammerer et al., 2009).

Despite these advantages, the mechanism of both tag-based and faceted search interfaces are based on reactive user feedback, not an interactive recommendation. The facet structures are pre-defined, either manually (Yee et al., 2003), computationally from document data (Dash et al., 2008; Kong & Allan, 2014), or the behavior of other users (Koren et al., 2008). Tag recommendations, on the other hand, are largely based on predicting relevant tags based on a user's previous selections. For example, the well-known Mr. Taggy system uses the tags that the user has activated and computes spreading activation to determine other tags being relevant given a particular combination of the tags and URLs that the user has selected (Kammerer et al., 2009). As a result, manually curated faceted navigation structures and tags inferred from initial search results are known to be well received by users and can improve search performance, but they rely either on manual faceted classifications or tags that are already available within the top-ranked search results. In the case of faceted search, this involves laborious manual design for every domain and application separately. In the case of tag recommendation, this potentially locks the user in the initial query scope. Thus, the user interactions are limited in selecting pre-computed facets or tags that are predicted to be relevant based on previous filtering activity without reflecting active feedback from the user.

In addition, users are often offered a fixed number of facets or tags to select from, usually just enough to fit on a screen. This can lead to the most effective facets not being offered even in the case when they are organized in hierarchical structures that can be navigated. In fact, on many information-intensive domains, the total number of possible facets or tags can be much larger than the number of facets that can be meaningfully presented for the user, sometimes by orders of magnitude. For example, a product or media data archive can consist of millions of entities that again are annotated with tens of thousands of different tags, each being a potential search parameter. Therefore, presenting a simple top-ranked set of filtering options for the user may not lead to interaction options best suited for a user's needs.

To address these shortcomings, we introduce *active tag recommendation*: a method in which the search system actively recommends tags that are predicted to be relevant for the user based on all previous interactions with the tag recommendation system. As a result, the system and the user can jointly reduce uncertainty to predict tags that are useful for the user. Consequently, active tag recommendation overcomes the problems of pre-defined "one size fits all" navigation structures and simple ranking approaches that are based on the activation of filters. Active tag recommendation can predict useful tags to be used as filters based on user active feedback without requiring the user to commit to activate tags or select them only from the currently high-ranked results.

1.1. Research objectives

The objective of our research is to investigate the effects of active recommendation of tags on users' interaction effectiveness and the resulting retrieval performance in entity search. We present a computational method and a system implementation of active tag recommendation. Using the system implementation, we empirically study the performance of the method as part of a fully functional experimental search system, in which the user can search using conventional interactions (Fig. 1A) and interactions via the tags recommended by the system (Fig. 1C).

We report results from a user experiment where active tag recommendation is compared to a control condition. A control condition represents an entity search system that allows users to issue written queries via a search box. Here, we refer to the interaction enabled in the control condition as written queries. The control condition also has a query autocompletion support for the written queries (Fig. 1A), but the active tag recommendation is not available in the control condition. In the study, participants conducted search tasks on a large real-world movie review dataset. Our empirical research objective is to study whether active tag recommendation can improve interaction with the search engine, improve the ranking of relevant information that ends up being selected by the user to be useful for their task, and affect search task completion time. In particular, we seek answers to the following research questions:

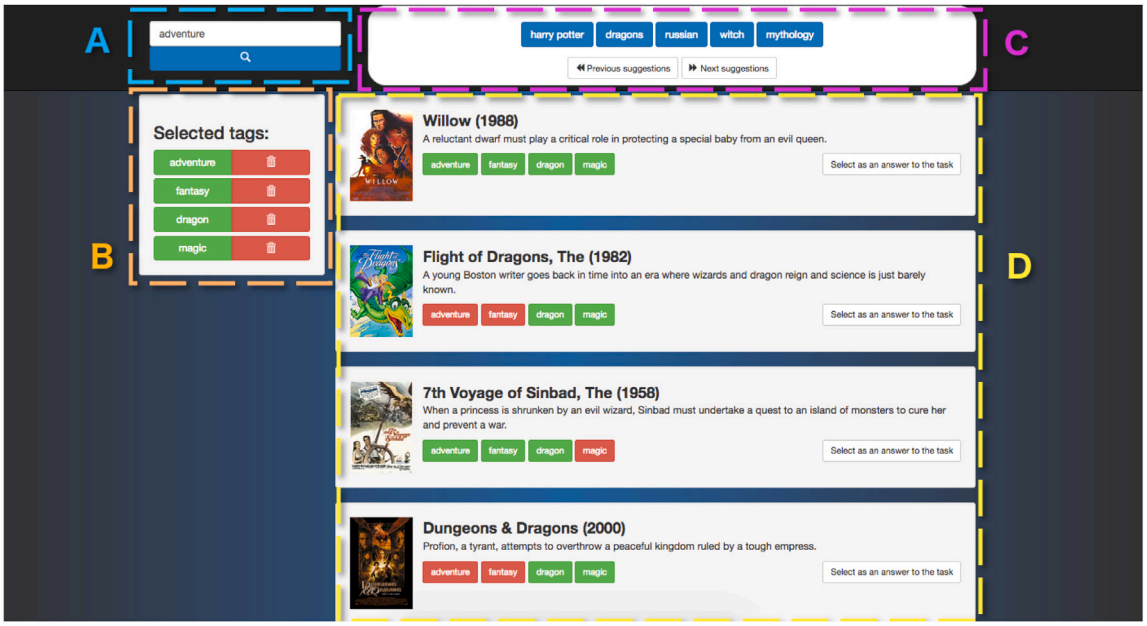

Fig. 1. The user interface of the experimental system. The user interface is composed of four elements. (A) the written query element. (B) The list of selected tags. (C) The list of tags that are actively recommended to the user. (D) The ranked entities with matching and non-matching tags are shown under each entity.

RQ1 Is active tag recommendation associated with improved search result ranking at a task level when compared to the control condition?

RQ2 Does active tag recommendation reduce or substitute written queries to direct search when compared to the control condition?

RQ3 Does active tag recommendation affect search task execution time when compared to the control condition?

1.2. Contributions and summary of findings

The theoretical and practical significance of the work arises from the computational method and system implementation for active recommendation for interactive entity search. The empirical findings from the analysis of search and interaction logs show that participants interact more and more effectively by using active tag recommendation over the control condition. Our results also show that active tag recommendation lead to an improved ranking of entities that participants' select to be useful to fulfill their tasks without compromising task completion time. In summary, the technical contributions and findings from the user experiment can be summarized as follows:

1. We present active tag recommendation — a technique that actively learns to recommend tags by exploring and exploiting interactive user feedback on the recommended tags. As opposed to fixed facet structures or exploitation of passive user behavior, active tag recommendation relies on explicit feedback between the recommended tags and the user.
2. We present a practical system implementation of active tag recommendation as a part of a functional entity search system indexed with a large collection of real-world movie data.
3. We show that active tag recommendation enables effective interaction with the search system leading to improved entity ranking in realistic task-based entity search.
4. We show that active tag recommendation allows effective interactions, but written query interaction remains important and is not substituted but complemented with active tag recommendation.
5. We show that active tag recommendation does not compromise task execution time despite the additional user interface elements and interaction options.

2. Background

Our work is related to several interleaved research areas ranging from entity-oriented search, task-based search, faceted and tag-based search to interactive search support and personalization in complex tasks. These are briefly reviewed in the following subsections.

2.1. Entity-oriented search

The entity-oriented approach to search relies on organizing and accessing information around entities and their relationships. Entities are often defined as the natural units to represent information as humans mostly think – and search information – in terms of real-world things and their connections to each other (Balog, 2018). Allowing users to interact with specific entities offers an effective alternative to conventional document-based information access. For example, representing a large movie archive as entities and their associations with other entities and tags via relationships to topics allows users to focus their search activities interactively on tags and other structured data representing meaningful search concepts. This provides a natural way to express information needs to filter the movies matching users' interests as opposed to keyword search that requires the user to come up with exact queries describing their information needs upfront.

Entity search has been recently studied widely in industry and academia. Recently, entity displays have become an integral part of commercial search engine result pages and their performance for effectiveness and user experience has been studied (Balog, 2018; Blanco et al., 2013; Chen et al., 2016; Gerritse et al., 2020; Hasibi et al., 2017; Huang et al., 2020; Reinanda et al., 2015). A user searching information related to entities, such as people, movies, music, or other information available in entity collections or knowledge-graphs is offered information about an entity or a set of entities directly on the results page, helping the user to find and navigate to the information without clicking separate links to documents (Balog, 2018). Such approaches can be seen as a form of entity recommendation in response to an individual written query (Blanco et al., 2013).

Another area of interest has been the task of ranking or selecting the most important facts about an entity and presenting it to the user as a coherent but compressed entity card (Hasibi et al., 2017). This problem has been widely investigated in various entity summarization tasks (Cheng et al., 2011; Gunaratna et al., 2015; Liu et al., 2021). A common approach has been to compute graph centrality on the knowledge graph to select the most important facts to be presented (Gerritse et al., 2020; Reinanda et al., 2015) or to compute factual ranking dynamically to summarize entity contents (Hasibi et al., 2017).

Less prominent but highly important research direction has been the presentation of entity information and personalized recommendations for users and enabling user interaction over entity data, such that it can be shown to benefit users in realistic search and interactive recommendation scenarios (Shi et al., 2015; Yu et al., 2014; Zhou et al., 2020). Recent work has also studied the problem in realistic user experiments with personally recorded entity data from in-situ data acquisition (Jacucci et al., 2021).

This line of research has suggested that entity recommendations can enhance user experience and assist users to accomplish their information needs interactively in dialog with the recommender system (Lei, Zhang et al., 2020), and even to offer a more comprehensive view to data in response to visualizing and exposing the entity data for interaction (Jacucci et al., 2021). While most of the work in entity search has been devoted for studying an individual query-response, recent research marks the trend toward entity recommendation that is interactive and can learn user needs and preferences on-line.

2.2. Task-based search

Task-based search refers to information-seeking activity that spans through a search session or even several sessions, consisting of sequences of queries and other interactions situated in a task context (Ingwersen & Järvelin, 2005). While this is the way many information searches are conducted, the majority of information retrieval research has been devoted to understanding relatively short search sessions and relied on interactions recorded from conventional search user interfaces (Carterette et al., 2016; Verma et al., 2016). During the last decade, increasing attention has been devoted to exploratory scenarios in which the success of search activity is often measured at the task level (Shah & White, 2021). However, the complex relationship between task-level goals, search behavior, and interactive system support are brought forward only very recently (Sarkar & Shah, 2021). The research community is still struggling to develop computational methods for supporting task-based search, and it has turned out challenging to move evaluation practices beyond simple analyses of single-query or single-turn interactions (Shah & White, 2021). Nevertheless, researchers have recognized the importance of session-level and task-level search support and methods using off-line log-data analyses have been introduced (Guan et al., 2013; He et al., 2013; Raman et al., 2014).

As such, the majority of previous work on session search and task-based search relies on simulation studies without addressing the real interactions between the search systems and their users when they are offered as a part of a search user interface. For example, session-based search has been studied by using the data from the Session Track at TREC (Carterette et al., 2016) and task-based search using the data from the Tasks Track at TREC (Verma et al., 2016). A disadvantage of these TREC tasks, however, is that they are based on conventional assumptions of the search user interface and user input available for the search system, as well as focus on ranked lists of documents instead of entities. These assumptions limit the advances that can be made by focusing on more structured entity data and user interface design. Thus, the roles of different end-user functionalities and interface techniques that may support users' tasks beyond written queries, as well as methods to evaluate these in interactive settings, are not extensively studied using TREC data. Thus, it remains unclear whether and how the findings resulting from analysis of log data recorded using standard search user interfaces can inform us about information-seeking performance with alternative search user interfaces.

2.3. Interactive search support

Faceted search (Yee et al., 2003) is one of the most deployed techniques to offer functionality for navigating complex information spaces. The advantage of faceted search is that by visualizing facets it provides guidance to understand and control to iteratively filter search results. The key idea behind faceted search is that it enhances a conventional search user interface with a faceted navigation component. The component provides the users with affordances to narrow down search results by applying multiple facets, often implemented as boolean filters, based on faceted classification. A faceted classification system defines a set of facets along multiple dimensions over the searchable information space. At search time, facets are activated and deactivated to enable filtering or re-ranking of the search results. Facets align with the properties of the data, such that they can be pre-existing fields in a database such as price or product type, or a manually curated classification system.

Written query search requires that the user constructs a query from scratch to represent the user's information need. This process may be supported by autocompletion or query suggestion. Yet, it is based on the user's active input to initialize and scope the query. In contrast, faceted search allows the user to activate facets to filter results in interaction with the system. The underlying faceted classifications are often based on data that are independent of the query. Moreover, the visualization of the facets can also reveal the internal structure of the data collection and help the user to comprehend the search space (Kules et al., 2009; Peltonen et al., 2017).

Faceted search tools have been evaluated in various information retrieval tasks. Manually constructed faceted classifications have been found effective in task-based user studies (Koren et al., 2008; Ruotsalo et al., 2020; Yee et al., 2003) and have also become standard elements in commercial systems. However, they require manually curated faceted classifications and are not directly usable for new domains and datasets without laborious customization.

Similar but more data-driven approaches to providing controls for interactive search are navigation support enabled by search space clustering (Hearst, 1995), visualization support for result comprehension and re-ranking (Matejka et al., 2012; Peltonen et al., 2017; Rahdari et al., 2020), visual re-ranking (Klouche et al., 2017), and information generation (Ukkonen et al., 2020). While these approaches provide means for visualizing search results and interacting with the underlying information space, they are primarily aiming to enhance information comprehension or sense-making rather than search effectiveness.

Another line of research has focused on directly supporting query construction on existing search interfaces in which the queries are expressed by writing them. These include, for example, query auto-completion (Jiang, Ke et al., 2014), query recommendation (Boldi et al., 2008), intent-driven search result re-ranking (Hu et al., 2011), and online entity recommendation reflecting a user's search behavior (Jacucci et al., 2021).

Query auto-completion provides additional query terms or entire queries to replace the initial query. It is effective for various search tasks (Bar-Yossef & Kraus, 2011; Shokouhi, 2013), in particular for complex search tasks that are not easily solvable with a single query (Capra et al., 2015). Query auto-completion and often visualization, however, are limited to assisting the user to construct queries that are already in the scope of the initial user input rather than suggesting the user with queries or terms for exploration. Queries can be predicted by implicitly monitoring the users' pre-search context (Kong et al., 2015) or by identifying alternative queries that can be inferred based on the original query and the relationships of the query and the user's longer-term behavioral history (Sordoni et al., 2015; Vuong et al., 2021). The suggested queries may therefore be good alternatives to the initial query or predicting the next query but are not necessarily helpful in exploring outside the initial query scope.

More empirically oriented research has also revealed that search support is used differently depending on the task and the phase of the search session (Jiang, He et al., 2014). This suggests that longer search sessions are associated with complex information needs, which are then reflected in increased user effort and the number of interactions required to obtain satisfactory results. Investigations of task-level search performance have revealed that when users are also examining results that are ranked lower, but that may still be relevant, they achieve improved task outcomes for more complex tasks (Vakkari & Huuskonen, 2012). Thus, focusing on system features that only optimize ranking in response to an individual query may falsely focus on the quality of the top-ranked results, while users may rather benefit from novel interactive features in search user interfaces to better explore the result space.

2.4. Dynamic session models

Session-level modeling has shown promise to increase the user's exposure to broader exploratory results or more scoped personalized results at a session level (Liu et al., 2020; Raman et al., 2013). Research has also found that tasks can often span across several sessions, and users may need varying support to orient and re-acquaint their search goals (Li et al., 2020). Therefore, online prediction of the support that users need for drifting search intentions has emerged as an important research direction (Andolina et al., 2015; Cheng et al., 2010; Mitsui et al., 2017; Ruotsalo et al., 2018; Zamani, Mitra et al., 2020). More recent research has also proposed dialogue techniques to interactively guide users to find the intended information via conversational search and recommendation approaches (Christakopoulou et al., 2016; Iovine et al., 2021; Zamani, Dumais et al., 2020; Zhang et al., 2018).

Researches have also incorporated session models more tightly into interactive search support. Dynamic faceted search (Dash et al., 2008) proposes an approach that selects a small set of "interesting" attributes to be presented for the user given an interestingness measure that captures relevance and surprisingness. Another dynamic approach is proposed to personalize interactive filtering by inferring filtering options from collaborative behavioral data (Koren et al., 2008). Here, the search interface is customized to each user's interaction behavior by associating it with other users' behavior. However, these approaches are not evaluated in user studies, but their theoretical performance is investigated only in simulations. Consequently, it remains unclear how the interestingness measure or collaborative filtering would reflect users' interaction behavior in a more realistic interactive search.

Table 1
Mathematical notation of the multi-armed bandit model used in active learning of tag recommendations.

Symbol	Explanation
T	Total number of iterations in the search session
t	Index of an iteration
F	Number of tags that have received feedback
r	Relevance feedback value for a particular tag
$\tilde{r}$	Vector of all F relevance feedback values received
K	Data matrix for tags with feedback
k	Vector for a tag whose relevance is to be predicted
λ	Regularization parameter in the regression model
a	Regression weight vector for a tag
$\hat{v}$	Relevance score for a tag (upper confidence bound)
c	Constant to balance exploration and exploitation
i	Index of a tag

Recently, the Estimation – Action – Reflection framework was proposed to integrate users' behavioral choices of recommended interaction options with their responses to these options (Lei, He et al., 2020). This framework combines conversational component and recommendation component and decomposes the recommendation task into three subproblems: what to ask, when to recommend, and how to adapt with user feedback. Our approach is different as it recommends tags at every iteration as part of a search process and does not maintain a conversational dialogue. It is similar in that it predicts the tags that are likely to be most useful for the user in a certain situation and elicits feedback from the users to update the recommender model when a user rejects or accepts the recommendations made by the model.

All in all, such studies show increased focus on utilizing user signals, whether explicit or implicit, to clarify users' information needs, decisions, and intentions. Our approach differs from previous work by promoting an active learning approach from user interactions with the tag recommendation component as a source for feedback. Active tag recommendation allows the user to provide feedback for the tag recommendation by selecting or ignoring the recommended tags without interleaving with the primary search activity.

3. Active tag recommendation

Active tag recommendation is formulated as an interactive learning problem that actively exposes recommendations for user interaction, which is then used as feedback to learn improved recommendations in the subsequent iterations. The tags are predicted by exploring and exploiting user feedback and the associated entities are ranked using the selected tags as the search session progresses.

3.1. Model

The model relies on entity-tag representation. Tags are associated with the entities indexed in the system. Using such representation, active tag recommendation utilizes an upper-confidence bound contextual multi-armed bandit model (Auer, 2003) to predict entities that are relevant for the user. Similar models have been successfully used for online learning in information retrieval (Ruotsalo et al., 2018; Tang et al., 2015), but here we apply them for online tag recommendation.

More precisely, interaction with the active tag recommendation can be formulated as a reinforcement learning problem with multi-armed bandits where the feedback for the tag prediction is *evaluative*, i.e., the feedback is not required to be optimal, and the system not only *exploits* to find the best tag recommendations given the selections so far but also *explores* tags that could be relevant but of which the user is not yet aware (Auer, 2003). This is in contrast to a conventional interaction element or query suggestion prediction problem where any observed historical interaction is often assumed to be optimal feedback and used directly to learn improved estimates. Therefore, the suboptimality of the selected tags has an important implication for the assumptions made for the interactions expected from the user: our approach does not require the user to make consistent selections or to commit to an interaction that is assumed to take the user closer to the intended target information, but it can learn from suboptimal feedback resulting from exploratory user behavior.

Intuitively, exploitation accounts for the expected value of a regression model (offers maximally relevant tags given the observed interactions), and exploration accounts for the uncertainty of the expected value (offers tags that are relevant but also uncertain). Accounting for exploration and exploitation simultaneously results in tag recommendations that are maximally relevant but have high uncertainty. In interaction with the system, the user can then reduce the uncertainty by rewarding the tags (arms in the multi-armed bandit model) that are found relevant for the task. By doing so, some of the tags increase their expected values, but some also increase their uncertainty. As the tags are ranked by balancing both the expected value and uncertainty, the model continues to explore different tags.

The formal notation for the model is given in Table 1. The data of tags and entities are stored in matrix K. That is, the vectors contain information on occurrences of tags over the entities. A tag recommendation session consists of $t = 1, \ldots, T$ iterations. For the model, user interactions are assumed to take the following form. At each iteration t, the top-k ranked tags are visualized for the user,

and the user either selects a tag or does not select a tag. Selecting a tag is interpreted as relevance feedback. Up to time t we have collected F_t instances of relevance feedback (selected tags) $r_1 \dots r_{F_t}$. The model also accounts for the tags that are recommended for the user, but not selected. These tags are known by the system. That is, they appear in an entity retrieved by the system, but they are not selected by the user. These tags and their connections to entities are stored in a submatrix K_t.

At each iteration, the model results in estimated relevance scores $\hat{v}_{t,1}, \dots, \hat{v}_{t,M}$ for all M tags. The estimated relevance scores are the upper bounds of the expected relevance scores. That is, they are both expected to be maximally relevant and maximally uncertain. The relevance scores are then used to rank the tags to be presented for the user at the next iteration.

The method then consists of two steps. First, it computes a regression weight for each tag using the context vectors containing the occurrences of the tags across entities seen so far stored in K_t. That is, for each tag $i = 1, \dots, M$ we denote the corresponding row of K by k_i and compute the weight vector

$$a_i = k_i(K_t^\top K_t + \lambda I)^{-1} K_t^\top, \tag{1}$$

Then, for each tag $i = 1, \dots, M$, the relevance score $\hat{v}_{t,i}$ is computed by incorporating the feedback obtained up to iteration t:

$$\hat{v}_{t,i} = a_i \cdot r_f + \frac{c}{2}\|a_i\|, \tag{2}$$

such that a ranking by $v_{t,i}$ corresponds to the most relevant tags to be visualized for the user in the next iteration.

The model yields the upper-confidence bound for each tag. This means that the relevance score reflects tags that are most relevant given the observed feedback (highest expected value), while at the same time being most uncertain for the model (highest uncertainty). This differs from the conventional idea of relevance estimation that maximizes only the expected value as it simultaneously maximizes both the model perspective for the optimal future feedback (maximal uncertainty) and the user perspective for maximal present relevance (maximal expected value).

3.2. User interface and interaction design

The model was implemented as a part of a search system that enables evaluating the approach as part of a fully functional entity search engine on a movie domain. The system also allows isolating different user interface components and thus studying their effect on user's performance and interactions as part of realistic information-seeking processes. In detail, the system was implemented to support various experimental conditions such that interactions with different components could be enabled and disabled dynamically across and within experimental conditions.

The interface of the system is composed of four main components, which are shown in Fig. 1. They consist of: a written-query element (area A), a list of selected tags (area B), a list of active tags recommended to the user (area C), and a list of search results (area D).

The written-query element allows the user to write queries and provides autocompleted query suggestions as the user types. The autocompletion matches written prefixes with the tags that are used to index the entities in the system. To submit a new query, the user can type and select a search tag from the suggested tags or write a complete query. The tag is then added to the list of selected tags, which represent the query. The list of selected tags shows a set of presently selected tags within the search session. At any time, the user can also remove tags from the list.

The list of actively recommended tags is shown on top of the screen (area C) after the initial query tag is selected using the written-query element. At each iteration, five active tag recommedńdations are shown along with "next suggestions" and "previous suggestions" when appropriate. The ‘‘next suggestions’’ button can be used to reveal the recommended tags beyond the five top-ranked tags that are visualized at the time. The ‘‘previous suggestions’’ button can be used to move back to the previous set of five tags. The user can select tags from the active tag recommendations by clicking them. Selecting an active tag results in adding the tag as a new query term (area B). In response to adding a tag, a new query is executed to retrieve new results and update all elements on the screen. The user can also remove a tag from area B to adjust the query. The list of search results (area D) is updated in response to adding or removing a tag. The resulting entities are shown as a ranked list in the result area D, along with the matching and non-matching tags under each entity.

Fig. 2 illustrates an interaction sequence when using the system. In step 1, the user decides to look for a movie involving ‘‘dystopia’’. The user uses the written-query element to search for this tag. A query auto-completion helps the user select an intended tag. In step 2, the active tag element suggests a list of related tags. The user decides to refine the search by selecting a tag from the list. In this case, the user adds the tag ‘‘cyberpunk’’. In step 3, the active tag element is updated, and new tags are recommended based on the previously selected tags. Search results are also updated in response to interactions with the written-query or active tag elements. Steps 2 and 3 can be repeated until the user is satisfied with the search results. In step 4, the user selects a movie.

4. User experiment

An experiment comparing active tag recommendation to a control system without active tag recommendation was conducted. We used an experimental design following the ‘‘All Other Things Being Equal’’ principle in which the experimental system with active tag recommendation was compared to a control condition that was the same as the experimental system except that the active tag recommendation component was removed.

Table 2
Task descriptions used in the experiments.

Task description
Select 5 movies that you would want to watch to keep you awake on a long flight.
Select 5 movies that contain romance and sci-fi, but avoid gun violence.
Select 5 movies that are funny and insightful.
Select 5 movies that you would watch with a grandparent or older relative.
Select 5 movies that you would watch if you wanted to avoid Hollywood movies.
Select 5 movies that are scary and futuristic.

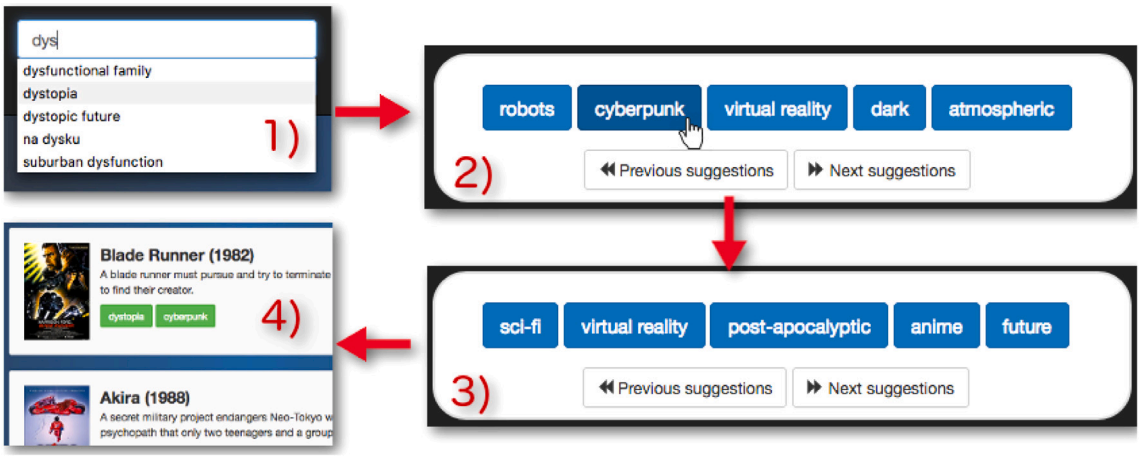

Fig. 2. A step-by-step illustration of a user running through the system. (1) The user searches for "dystopia". (2) Active tag recommendations are updated. The user clicks on "cyberpunk". (3) Active tag recommendations are updated again. (4) Results are shown.

4.1. Experimental design

A within-subjects experiment was designed with two system configurations: active tag recommendation and a control condition. The system and task order were counterbalanced, but all conditions were used by each participant to exclude confounding factors originating from user variance.

4.2. Control condition

The control condition was the same system, but the active tag recommendation element itself was removed. All other things were held constant in both conditions. Both systems used the same ranking model, the same query autocompletion model, and were indexed with the same data collection. In summary, the user interface was the same as in the experimental condition, but the active tag recommendation component (area C), shown on the top of Fig. 1, was removed. The interface still presented the selected tags chosen by the user (area B) via the query autocompletion (area A).

4.3. Participants

A total of 60 participants were recruited via email lists and social media services. Valid logs were captured for 45 participants whose data was used in the analysis. The rest of the participants had technical problems with the logged data due to accessing the system with an improper client (mobile or otherwise non-compliant browser), or they left the experiment partway through. The median age of the participants with valid data was 24. The participants reported their gender to be female (19), male (26), and other (0). Three participants reported vocational school as their highest level of education, 13 reported college, 15 reported bachelor's degree, four reported master's degree, and one reported Ph.D., and the rest reported other. The participants reported themselves mentally and physically healthy, and they reported that they had normal vision, including normal color vision.

4.4. Procedure

The experiment was conducted online. Participants received a link to the instructions of the experiment and they were first presented with an information screen that explained the purpose of the data usage only for scientific purposes. The screen also provided information that the participation was voluntary and quitting the experiment would not have any negative consequences for the participants. Then, the system presented another instruction screen, which explained all features of the system and how they

could be used. The participants were then given a test task to try out the system. The data from the test task was not used in further analysis. After the instructions, the participant was directed to the actual system.

Then the actual experiment started. The participants were then given a series of tasks to complete by using the system. Each task started with an instruction screen. Before each task, the participants confirmed that they understood the task and were ready to conduct the task without interruptions. The configurations of the systems and the tasks were counterbalanced automatically in the backend so that every participant completed six tasks (three tasks with each system configuration) in varying, counterbalanced order, such that every new participant entering the experiment started with a different task and a different system than the previous participant. The participants could freely choose how to use the offered system features, and they were not forced or encouraged to use any particular features. The participant was shown a timer in the corner of the screen and notified that after 10 min the task would be terminated, but the participants were not encouraged to complete the tasks faster than they would normally do. All participants completed all six tasks within 5 min (300 s), well before the time limit.

4.5. Data collection, indexing, and ranking

The systems were indexed with the MovieLens 20M dataset[1] which includes 465,564 user-submitted tags for 27,278 movies released between 1995 and 2015. The search engine used only the tags as the descriptors for the movies to allow comparability between the active tag recommendation and conventional searching via written queries and autocompleted tags. The tags occurring less than five times were removed, resulting in over 3000 unique tags. A straightforward vector space model with tf–idf weighting and cosine similarity was used to rank the movies in response to the selected tags.

4.6. Search tasks

The participants were requested to conduct search tasks in the movie domain. The tasks were complex enough to ensure that exploration is necessary for participants to select the information to accomplish the task. This followed the task design strategy commonly adopted for studying exploratory search systems (Kules & Capra, 2009; Wildemuth & Freund, 2012). The tasks were designed in a way that participants could not solve them using a single interaction. Thus, allowing participants to choose the kind of interactions that best supported the tasks.

Search tasks were presented using a general instruction: You are asked to compose a list of candidate movies to watch in a given scenario. You are provided with a series of scenarios and a search system to select the movies in each scenario. You have 10 min to complete searching the movies in each scenario.

Search tasks were specified for each scenario and are shown in Table 2. They consisted of finding five movies that fulfill a set of qualifications as specified by the task description. The qualifications were selected in such a way that they would not have a direct match to tags and would require the user to engage with the system.

The participants were free to use the provided systems as they would use any search system. To complete the given tasks, the participants collected five movie entities during each task by adding them to the set of entities that they wished to submit as an answer. This was performed by clicking the "select as an answer to the task" button appearing next to each entity in the search result listing.

4.7. Data collection and logging

The experiments were run online, and participants conducted the experiments using their own devices. To avoid any dependencies on the client-side software, a comprehensive logging system was implemented in the backend. The logging system captured task information and anonymous user identification via cookies. The identification ensured that each user was only able to take part in the experiment once. The logging system also recorded interaction data and information retrieved in response to the interactions, including written queries, retrieved and selected entities and tags, and time stamps of the occurrence of each interaction. The positions of all entities were logged at each iteration to be able to track the effect of interactions on the rank positions of the selected entities.

After the data collection, we inspected that the information selected made sense in each task scenario. This was performed to ensure that the participants had conducted the tasks properly. All of the participants who completed all of the tasks had data that was appropriate and their data were included in the final analyses.

4.8. Measures

The user experiment design allowed collection of rich interaction data on participants' interaction behavior at task-level and the corresponding effects on entity ranking. These data were used to define three types of measures. First, measures that reflect the ranking effectiveness of the selected tags in response to user interactions. Second, a measure to quantify interaction effectiveness to obtain the results that are selected by the users. Third, a measure of time duration to complete the given tasks.

[1] https://grouplens.org/datasets/movielens/20m/.

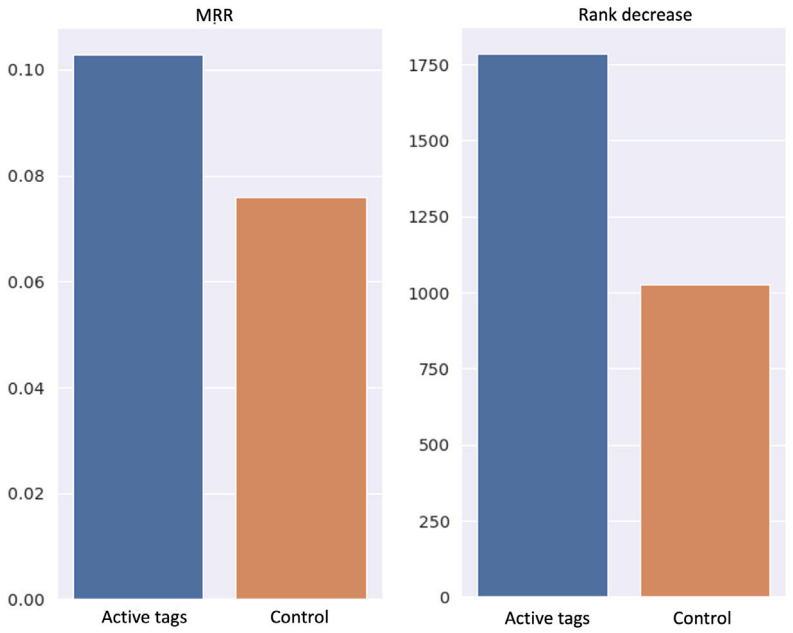

Fig. 3. The results of search effectiveness measures for the compared conditions averaged over the users and tasks. Left: The selection Mean Reciprocal Rank (MRR). Right: The mean decrease of rank.

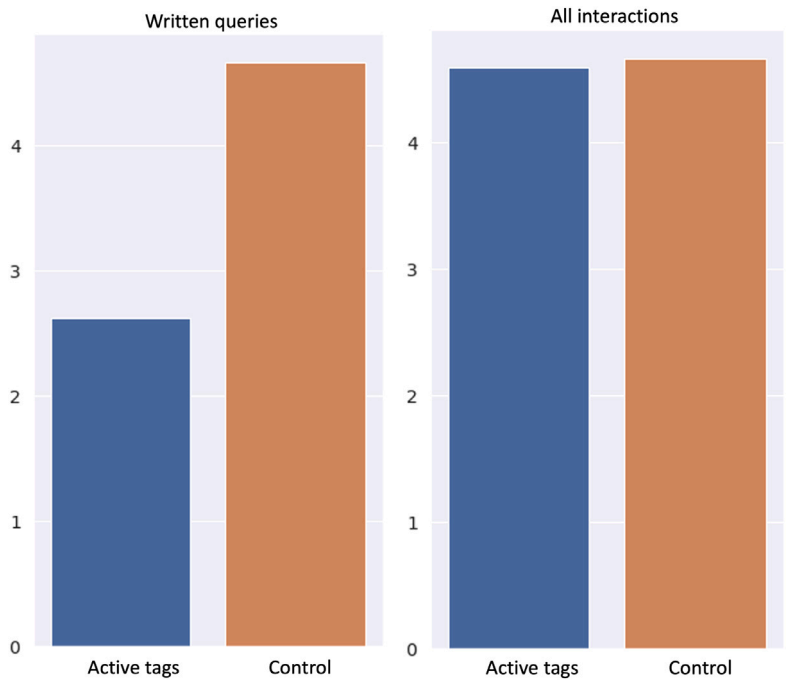

Fig. 4. The results of interaction effectiveness measures for the compared conditions averaged over the users and tasks. Left: The number of written queries. Right: The number of all interactions (including both written queries and active tag recommendations in the active tag recommendation condition).

Ranking performance in response to interaction

As the participants' tasks were information selection tasks where they were asked to choose results that matched the tasks within a limited time, we selected ranking measures that rely on the ranking position of a known target entity that was at the end selected by the participant: mean reciprocal rank (MRR) of selected information (Craswell, 2009). MRR is a suitable measure in situations

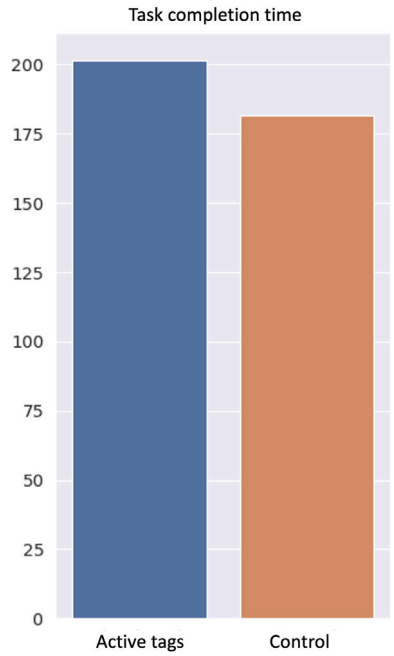

Fig. 5. Mean task completion time in seconds averaged over the users and tasks for the compared conditions.

where the focus is on the rank of a selected target entity, rather than the ranking of the entities in the entire search engine response. In our case, the target entities are the entities selected by the participant to fulfill the tasks.

Intuitively, in our setting, the MRR measure reflects how effective a specific type of system condition is in decreasing the rank of the entity that the user selects as the outcome of the task (i.e. better ranked results are in smaller ranks and higher in a ranked list). The higher the MRR, the better the particular interactive system feature tested in the condition is associated with the decreased rank of the target entities. Thus, MRR reflects the utility of the interactive system feature for ranking, and in turn, the utility of the system feature present in the condition for finding the selected entity.

Interaction effectiveness

Interaction effectiveness was measured via two measures. The first measure quantifies the frequency of interactions. Interactions are accounted separately for written queries and active tag recommendation selections. We also quantified the total number of interactions performed to conduct or adjust a query (i.e. the total frequency of written queries and actively recommended tag selections).

Second, we measured the mean decrease of rank. The mean decrease of rank quantifies the effect of a decrease of rank in response to a certain type of user interaction (i.e. a beneficial interaction should make the ranks smaller and the target results to appear higher in a ranked list). While ranking performance (MRR) measures an average ranking for the entire system condition, a mean decrease of rank measures the average ranking performance of an interactive element. Again, we measure two types of user interactions: written queries and activating a recommended tag. Intuitively, the relative decrease of the rank of selected entities with respect to an interaction element measures whether interactions with that element were associated with the selected entities to be ranked higher than they were when different interactions were performed. If users interact more with active tag recommendations, and those interactions lead the selected movies to decrease their ranks more than the other interactions, then interaction with the recommended tags is associated with higher effectiveness than the other interaction options. This measure was only computed in the active tag recommendation condition as that was the condition that allowed the comparison of the interaction elements when both of the elements were available for the users.

Task completion time

Task completion time was measured as the duration from the beginning of the task to the point when the user had successfully selected the five movies meeting the task criteria.

5. Results

An overview of the results of the experiments is shown in Table 3 and illustrated in Figs. 3–5. Below, we will present the results with respect to each selected measure averaged over users and tasks.

Table 3
Experimental results. Statistically significant differences were found in selection MRR, rank decrease, and the number of written queries. Statistically significant differences were not found in task completion time and the number of all interactions. ± indicates the standard error of the mean.

Measure	Control	Active tag recommendation	p-value
Task completion time	181.7 s ± 132	201.3 s ± 176	$p = 0.453$
Selection MRR	0.76 ± 0.176	0.103 ± 0.210	$p < 0.001$
Rank decrease	1027 ± 3790	1783 ± 4192	$p = 0.004$
Number of written queries	4.66 ± 3.17	2.62 ± 2.032	$p < 0.001$
Number of all interactions	4.66 ± 3.17	4.59 ± 3.38	$p = 0.358$

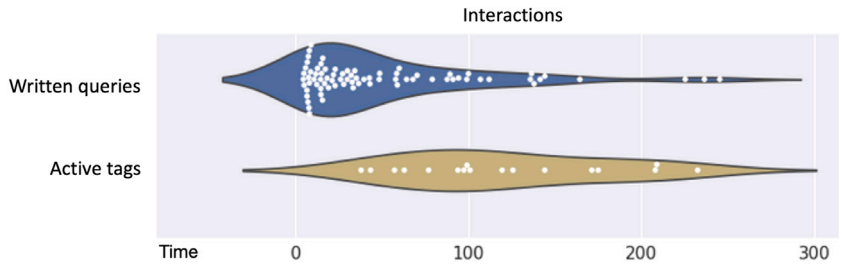

Fig. 6. Interaction distributions over time across the two interaction types: written queries and active tag recommendations. The plotted written-query interactions are more frequent as they include all letter-level written input to the written-query element. All tasks were completed within approximately 300 s.

5.1. Ranking performance

The ranking of selected information as measured by the MRR on the selected entity ranks was found to be higher in the active tag recommendation condition ($MRR = 0.103$) when compared to the control condition ($MRR = 0.76$). The difference was found to be significant according to Mann–Whitney–Wilcoxon Test, $p < 0.001$. The effect of interactions on the rank of the selected entities was also higher for the active tag recommendation than it was for the written-query interaction. The rank decrease for the active tag recommendation was on average 1783 positions, while it was 1027 positions for the written-query interaction. The difference was found to be significant according to Mann–Whitney–Wilcoxon Test, $p < 0.001$.

5.2. Interaction effectiveness

Interactions were found to be more effective for the active tag recommendation condition than they were for the control condition. The participants used similar amounts of recommended tags per task (1.97) than they used written queries (2.62) per task when presented with the system with active tag recommendation condition, in which both interaction options were available. The difference against the number of written queries used in the control condition (4.66) was found to be significant according to Mann–Whitney–Wilcoxon Test, $p < 0.001$. However, significant differences were not found (Mann–Whitney–Wilcoxon Signed Rank Test, $p = 0.358$) in the total number of interactions between the systems. There were on average (4.66) interactions in the control condition per task and ($\mu = 4.59$) interactions on average in the active tag recommendation condition per task.

The interactions had different temporal distributions, as shown in Fig. 6. Written queries were used more at the beginning of the task, while the active tag recommendations were used more throughout the task. This suggests that active tag recommendations were more likely used to specify the query after the initial search results were obtained through queries and query refinements.

5.3. Task completion time

Task completion time was found to be comparable between the systems. The participants in the control condition spent on average 181.7 s to complete the tasks and 201.3 s in the active tag recommendation condition. The differences between the task completion times were not found to be significant (Mann–Whitney–Wilcoxon Test, $p = 0.453$).

6. Discussion and conclusions

We introduced active tag recommendation and demonstrated it as a part of an interactive information retrieval system. The technique was evaluated in an online user experiment. Here, we summarize our contributions and findings, both methodological and empirical.

6.1. Answers to research questions

RQ1: Is active tag recommendation associated with improved search result ranking at a task level when compared to the control condition?

Yes, our results suggest that active tag recommendation improved the average ranking of the selected information at a task-leve. This was observed as an improved average MRR indicating an average higher rank of the selected information and as an improved mean rank in response to an interaction with the active tag recommendation element. This shows that while both, written queries and tag recommendations, were utilized by the participants, an average utility to decrease the rank of an entity that the participants ended up selecting as an answer to their task was significantly higher for the recommended tags. A possible explanation is that on a complex domain, such as movies, the entities are described with high-level tags, such as "scifi", but also with more specific tags, such as "cyberpunk". The users' ability to come up with a query that would distinguish the relevant movies associated with more general tags may be limited and might require several alternative queries and query refinements to come up with the correct search vocabulary.

RQ2:Does active tag recommendation reduce or substitute written queries to direct search when compared to the control condition?

Yes, our results suggest that active tag recommendation enables more effective interaction between the search engine and the user. The rank decreases of the selected information were significantly higher in the active tag recommendation condition than they were in the control condition. Participants also interacted more in the active tag recommendation condition, suggesting that active tag recommendation was perceived useful.

However, active tag recommendation did not replace the written-query interaction completely but rather augmented that interaction. This was observed despite the fact that query autocompletion suggestions were offered, indicating that in many cases, users preferred interaction with active tag recommendations over more conventional query autocompletion.

It is noteworthy that written-query interaction was still frequently used by the participants. According to the logs, it was the primary type of interaction at the beginning of the search session. However, the rank decrease of the selected entities resulting from interaction via written queries was significantly lower than it was for the actively recommended tags. On the other hand, the variance of decrease in rank was much lower for the written-query interactions. A possible explanation is that written queries are highly effective for initial expressions of information needs, but when they require further refinements and participants exhibit exploratory search behavior, active recommendation of tags can have high utility.

RQ3: Does active tag recommendation affect search task execution time when compared to the control condition?

No, active tag recommendation did not lead to shorter task execution times. The task execution time was found to be comparable to the control condition, and statistically significant differences were not found. On the other hand, participants also did not need to compromise task execution time, even when they had more interaction options available, and they interacted more. The results suggest that participants did not conduct tasks faster, but they were also not asked to do so. This is in line with user behavior revealed in previous research that has suggested that time may not be a good performance measure in exploratory search evaluation (Capra et al., 2007; White et al., 2008) and that task complexity plays a more important role in search satisfaction (Capra et al., 2015). This suggests that active tag recommendation did not make searching faster, but also did not cause additional user effort that would have been manifested in increased task durations.

In summary, the participants utilized active tag recommendation by substituting written-query interaction with tags recommended for them. This interaction engagement transferred to improved ranking without compromising the task completion time. All in all, our findings demonstrate the importance of designing for interactivity in supporting active user involvement in query specification and exploratory search. These designs provide opportunities to enable rich interactions and interfaces that can actively learn from the users.

6.2. Practical implications for designing interactive entity search assistance

The analysis allowed us to derive the following implications, which are of practical use for researchers and practitioners designing interfaces or interactive search support methods for entity search.

- Users choose to use multiple interactive elements to express their information needs in entity search and the proposed active tag recommendation is utilized by users substituting approximately half of the written queries. The users, however, do not increase the number of interactions they perform. This suggests that some form of active assistance to select search filters might be preferred by the users over recalling queries without any support.
- Active tag recommendations show more benefits than conventional written-query interaction for ranking the information that is selected by users. This suggests that the tags offered by the active tag recommendation may be more precise descriptors of the information need than the queries that the users write. On the other hand, the active tag recommendations are used more in the latter parts of the search sessions, which may indicate that written queries are used as initial, more general expressions of information needs, while the recommended tags are used to drill down to more specific entities under the general area.
- The time spent searching was not affected by the experimental condition. This suggests that active tag recommendation does not lead to additional exploration that might be performed simply because the users are exposed to an increased amount of options and stimulated with additional exploration options. This may indicate that richer user interfaces are beneficial even when they may increase the complexity of interaction and draw user attention during the interaction.

6.3. Contributions and their relation to previous work

Specialized search systems increasingly use interactive user interfaces to support their users in exploring large multi-dimensional collections of data; often stored and indexed as entities that are associated with some structured descriptors, for example, tags. This motivates research on understanding how auxiliary data, such as tags, used to describe the entities being searched can support search activities. Most previous approaches on entity search and tag recommendation rely on reactive user involvement, where the user is only expected to confirm filtering suggestions offered by the system, but the system does not actively expose options to learn from the user for improving the suggestions in the subsequent iterations. Therefore, these techniques are often seen as relying on passive support and do not actively engage the user in interaction with the recommendations.

The existing tag recommendation methods have exploited various information sources to predict the tags the user might want to use for searching, but they all rely on behavior log data that the user has presently or historically interacted or modeling the content data itself. Established tag recommendation approaches have relied on exploiting tag-entity correlations with topic models (Krestel et al., 2009), tag co-occurrence structures modeled as graphs (Song et al., 2011), and comprehensive entity descriptions (Belém et al., 2014). Recently, neural models have been proposed for tag recommendation. For example, Zuo et al. (2016) proposed a representation learning approach to model the tag-entity connections using a lower-dimensional representation learned using a sparse autoencoder. A recent research unified many sources by using a hierarchical attention model (Sun et al., 2021). The approach simultaneously models content information, collaborative information, and personalized information. Consequently, the model integrates several information sources and learns to attend to the most relevant source for recommendations. There have been many follow-up studies in this direction with various deep learning approaches. For these, we refer the reader to a recent survey (Zhang et al., 2019).

Conversational search and recommender systems (Christakopoulou et al., 2016; Iovine et al., 2021; Zamani, Dumais et al., 2020; Zhang et al., 2018) come closest to our approach as they are designed to elicit preferences from their users interactively under an online learning framework. However, most of the studies so far are simulations of conversational dialogue and exploit log data collected off-line.

The present work is different as it approaches the tag recommendation problem as a type of online active learning. While the classic active learning aims at optimal acquirement of information from the oracle or user, the present work simultaneously explores and exploits the potential space of tags by using their associations in the entity data as context vectors. The user's interactions with the recommendations are used as feedback to guide the exploration and exploitation process.

To this end, active tag recommendation involves the user in the recommendation process by actively learning to recommend tags that might be interesting for the user to direct their search. The recommended tags are predicted from a large set of tags associated with the entities being searched via a multi-armed bandit model that accounts for the exploration/exploitation tradeoff of reinforcement learning (Auer, 2003). As a result, the system can expose the user with tags that are relevant to the user feedback (exploitation), but at the same time uncertain for the model (exploration).

In the present study, we investigated the effect of deployment of active tag recommendation as a part of a realistic search system that included other standard query construction features, such as query autocompletion. We aimed at identifying the effects of active tag recommendation for ranking effectiveness, interaction effectiveness, and time used for searching.

Previous research has mainly focused on computational evaluation. This is sound and widely applied evaluation methodology when the target is not to study real user interactions but may not reveal realistic user performance when used as part of an interactive system. That is, while the performance of the previously proposed methods is mainly assessed on existing datasets by predicting future interactions from previous logged interactions, our evaluation is not limited to scenarios where the interactions with the users are assumed to be accurately captured in logs or stay unaffected by user interface designs. Previous research is typically based on historical log data or simulations using and does not allow studying active involvement of the user or the effects of user interface designs on the recommendation process.

The present work focuses on studying the effects of active tag recommendation as part of a functional interactive entity search system. The present experiments emphasize empirical findings that result from task-based realistic interaction behavior. This allows quantifying the utility of active tag recommendation by using measures that focus on interaction effectiveness, ranking, as well as time spent completing search tasks. Furthermore, our experiments relied on a controlled experimental design where the same exact system being offered to the user without the active tag recommendation functionality. Our findings suggest that the users' search and information-selection behavior was affected by active tag recommendation without compromising task execution times. In particular, we showed that active tag recommendation can be an advantage to users for tasks that require exploring data beyond simple written queries.

More generally, our findings show that active tag recommendation is actually used and how it is used together with more conventional search interaction. Our findings also highlight that well-targeted active tag recommendations are more effective interactions to decrease the ranks of entities that the users' end up selecting than written queries. This makes our work unique from the empirical experimentation perspective and implies that user experimentation with novel entity search and recommendation is necessary to reveal the utility of interactive technology.

6.4. Limitations and future work

The findings from our experiments should be understood in the context of the experimental design that used a well-grounded control condition. The results revealed significant differences between the experimental condition and the control condition. However, in the present experiment, we did not compare active tag recommendation with other possible interactive interfaces, such as manually curated faceted classifications. Indeed, we believe that manually curated faceted search interfaces can lead to similar or even better interactions and search effectiveness as tag recommendations. However, such a system is dependent on manual curation conducted for every collection and domain separately and could not be considered an entirely fair comparison. Despite this, future work could quantify benefits over different types of interfaces and search support methods, such as alternative interfaces and methods for query suggestions, more advanced query auto-completion, or similar techniques that could be used as comparison conditions.

Future experiments could also study temporal effects or the need for search assistance in a situation when the searcher is struggling. For example, future experiments could include a condition where tags are recommended only after the first user query, or in some other limited regime through the search session, or in a situation when the user seems to be struggling with repeating reformulations of the initial query.

Our present methodological choices did not easily allow comparison of different computational approaches as all of the approaches would have to be tested as separate conditions in user experiments. Future work could explore the effect of different computational approaches for recommending tags. These could be driven by simulations and off-line comparisons of different methods using logged interaction data.

In the experiments, we limited our focus to specific tasks where users are primed to explore information and are likely to benefit from interactive system support. The task formulation may have led the participants to exhibit behavior that is in favor of our approach. Consequently, our findings may have limited utility in tasks that rely on repeated lookup search or that are easy to solve with simpler query autocompletion techniques as implemented also in the control condition. Moreover, we recruited participants and conducted experiments online. While this is in many ways a more realistic setting than pure laboratory experimentation, we do not have control over the participants' personal characteristics or pre-knowledge on the search domain. Future work could observe natural system usage by exposing the system to real users and their behavior when they search a variety of content with real-world information needs.

CRediT authorship contribution statement

Tuukka Ruotsalo: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. **Sean Weber:** Software, Visualization, Investigation. **Krzysztof Z. Gajos:** Supervision, Writing – review & editing.

Acknowledgment

The work was partially funded by the Academy of Finland.

References

Andolina, S., Klouche, K., Peltonen, J., Hoque, M., Ruotsalo, T., Cabral, D., Klami, A., Głowacka, D., Floréen, P., & Jacucci, G. (2015). IntentStreams: Smart parallel search streams for branching exploratory search. In *Iui '15, Proceedings of the 20th international conference on intelligent user interfaces* (pp. 300–305). New York, NY, USA: Association for Computing Machinery.

Auer, P. (2003). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research, 3*, 397–422.

Balog, K. (2018). *Entity-oriented search*. Springer Nature.

Balog, K., & Neumayer, R. (2013). A test collection for entity search in DBpedia. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 737–740).

Balog, K., Serdyukov, P., & de Vries, A. P. (2011). Overview of the TREC 2011 entity track.. In *TREC, Vol. 2011* (p. 11).

Bar-Yossef, Z., & Kraus, N. (2011). Context-sensitive query auto-completion. In *Www '11, Proceedings of the 20th international conference on world wide web* (pp. 107–116). New York, NY, USA: ACM.

Basu Roy, S., Wang, H., Das, G., Nambiar, U., & Mohania, M. (2008). Minimum-effort driven dynamic faceted search in structured databases. In *Proceedings of the 17th acm conference on information and knowledge management* (pp. 13–22). New York, NY, USA: ACM.

Belém, F. M., Martins, E. F., Almeida, J. M., & Gonçalves, M. A. (2014). Personalized and object-centered tag recommendation methods for web 2.0 applications. *Information Processing & Management, 50*(4), 524–553.

Blanco, R., Cambazoglu, B. B., Mika, P., & Torzec, N. (2013). Entity recommendations in web search. In *International semantic web conference* (pp. 33–48). Springer.

Boldi, P., Bonchi, F., Castillo, C., Donato, D., Gionis, A., & Vigna, S. (2008). The query-flow graph: Model and applications. In *Proceedings of the 17th acm conference on information and knowledge management* (pp. 609–618). New York, NY, USA: ACM.

Bota, H., Zhou, K., & Jose, J. M. (2016). Playing your cards right: The effect of entity cards on search behaviour and workload. In *Chiir '16, Proceedings of the 2016 acm on conference on human information interaction and retrieval* (pp. 131–140). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2854946.2854967.

Capra, R., Arguello, J., Crescenzi, A., & Vardell, E. (2015). Differences in the use of search assistance for tasks of varying complexity. In *Sigir '15, Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 23–32). New York, NY, USA: ACM.

Capra, R., Marchionini, G., Oh, J. S., Stutzman, F., & Zhang, Y. (2007). Effects of structure and interaction style on distinct search tasks. In *Jcdl '07, Proceedings of the 7th acm/ieee-cs joint conference on digital libraries* (pp. 442–451). New York, NY, USA: Association for Computing Machinery.

Carterette, B., Clough, P., Hall, M., Kanoulas, E., & Sanderson, M. (2016). Evaluating retrieval over sessions: The TREC session track 2011–2014. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 685–688).

Chen, J., Xiong, C., & Callan, J. (2016). An empirical study of learning to rank for entity search. In *Sigir '16, Proceedings of the 39th international acm sigir conference on research and development in information retrieval* (pp. 737–740). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2911451.2914725.

Cheng, Z., Gao, B., & Liu, T.-Y. (2010). Actively predicting diverse search intent from user browsing behaviors. In *Proceedings of the 19th international conference on world wide web* (pp. 221–230). New York, NY, USA: ACM.

Cheng, G., Tran, T., & Qu, Y. (2011). Relin: relatedness and informativeness-based centrality for entity summarization. In *International semantic web conference* (pp. 114–129). Springer.

Christakopoulou, K., Radlinski, F., & Hofmann, K. (2016). Towards conversational recommender systems. In *Kdd '16, Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 815–824). New York, NY, USA: Association for Computing Machinery.

Craswell, N. (2009). Mean reciprocal rank. In *Encyclopedia of database systems* (p. 1703). Springer.

Dash, D., Rao, J., Megiddo, N., Ailamaki, A., & Lohman, G. (2008). Dynamic faceted search for discovery-driven analysis. In *Cikm '08, Proceedings of the 17th acm conference on information and knowledge management* (pp. 3–12). New York, NY, USA: Association for Computing Machinery.

Dimitrov, D., Helic, D., & Strohmaier, M. (2018). Tag-based navigation and visualization. In *Social information access* (pp. 181–212). Springer.

Gerritse, E. J., Hasibi, F., & de Vries, A. P. (2020). Graph-embedding empowered entity retrieval. *Advances in Information Retrieval, 12035*, 97.

Guan, D., Zhang, S., & Yang, H. (2013). Utilizing query change for session search. In *Sigir '13, Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 453–462). New York, NY, USA: ACM.

Gunaratna, K., Thirunarayan, K., & Sheth, A. (2015). Faces: diversity-aware entity summarization using incremental hierarchical conceptual clustering. In *Twenty-ninth aaai conference on artificial intelligence*.

Hasibi, F., Balog, K., & Bratsberg, S. E. (2017). Dynamic factual summaries for entity cards. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval* (pp. 773–782).

He, J., Bron, M., & de Vries, A. P. (2013). Characterizing stages of a multi-session complex search task through direct and indirect query modifications. In *Sigir '13, Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 897–900). New York, NY, USA: ACM.

Hearst, M. A. (1995). TileBars: visualization of term distribution information in full text information access. In *Chi '95, Proceedings of the sigchi conference on human factors in computing systems* (pp. 59–66). New York, NY, USA: ACM Press/Addison-Wesley Publishing Co., ISBN: 0-201-84705-1.

Hu, B., Zhang, Y., Chen, W., Wang, G., & Yang, Q. (2011). Characterizing search intent diversity into click models. In *Proceedings of the 20th world wide web conference* (pp. 17–26). New York, NY, USA: ACM.

Huang, J., Wang, H., Zhang, W., & Liu, T. (2020). Multi-task learning for entity recommendation and document ranking in web search. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(5), 1–24.

Ingwersen, P., & Järvelin, K. (2005). *The turn: integration of information seeking and retrieval in context (the information retrieval series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc..

Iovine, A., Lops, P., Narducci, F., de Gemmis, M., & Semeraro, G. (2021). Improving preference elicitation in a conversational recommender system with active learning strategies. In *Sac '21, Proceedings of the 36th annual acm symposium on applied computing* (pp. 1375–1382). New York, NY, USA: Association for Computing Machinery.

Jacucci, G., Daee, P., Vuong, T., Andolina, S., Klouche, K., Sjöberg, M., Ruotsalo, T., & Kaski, S. (2021). Entity recommendation for everyday digital tasks. *ACM Transactions on Computational-Human Interaction*, *28*(5).

Jiang, J., He, D., & Allan, J. (2014). Searching, browsing, and clicking in a search session: Changes in user behavior by task and over time. In *Sigir '14, Proceedings of the 37th international acm sigir conference on research and development in information retrieval* (pp. 607–616). New York, NY, USA: ACM.

Jiang, J.-Y., Ke, Y.-Y., Chien, P.-Y., & Cheng, P.-J. (2014). Learning user reformulation behavior for query auto-completion. In *Sigir '14, Proceedings of the 37th international acm sigir conference on research and development in information retrieval* (pp. 445–454). New York, NY, USA: ACM.

Jones, R., & Klinkner, K. L. (2008). Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs. In *Proc. cikm '08* (pp. 699–708). New York, NY, USA: ACM.

Kammerer, Y., Nairn, R., Pirolli, P., & Chi, E. H. (2009). Signpost from the masses: Learning effects in an exploratory social tag search browser. In *Chi '09, Proceedings of the sigchi conference on human factors in computing systems* (pp. 625–634). New York, NY, USA: Association for Computing Machinery.

Klouche, K., Ruotsalo, T., Micallef, L., Andolina, S., & Jacucci, G. (2017). Visual re-ranking for multi-aspect information retrieval. In *Chiir '17, Proceedings of the 2017 conference on conference human information interaction and retrieval* (pp. 57–66). New York, NY, USA: ACM.

Kong, W., & Allan, J. (2014). Extending faceted search to the general web. In *Proceedings of the 23rd acm international on conference on information and knowledge management* (pp. 839–848). New York, NY, USA: ACM.

Kong, W., Li, R., Luo, J., Zhang, A., Chang, Y., & Allan, J. (2015). Predicting search intent based on pre-search context. In *Proceedings of the 38th international acm sigir conference on research & development in information retrieval* (pp. 503–512). New York, NY, USA: ACM.

Koren, J., Zhang, Y., & Liu, X. (2008). Personalized interactive faceted search. In *Www '08, Proceedings of the 17th international conference on world wide web* (pp. 477–486). New York, NY, USA: ACM.

Krestel, R., Fankhauser, P., & Nejdl, W. (2009). Latent dirichlet allocation for tag recommendation. In *Proceedings of the third acm conference on recommender systems* (pp. 61–68).

Kules, B., & Capra, R. (2009). Designing exploratory search tasks for user studies of information seeking support systems. In *Proceedings of the 9th acm/ieee-cs joint conference on digital libraries* (pp. 419–420).

Kules, B., Capra, R., Banta, M., & Sierra, T. (2009). What do exploratory searchers look at in a faceted search interface?. In *Proceedings of the 9th acm/ieee-cs joint conference on digital libraries* (pp. 313–322).

Lei, W., He, X., Miao, Y., Wu, Q., Hong, R., Kan, M.-Y., & Chua, T.-S. (2020). Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Wsdm '20, Proceedings of the 13th international conference on web search and data mining* (pp. 304–312). New York, NY, USA: Association for Computing Machinery.

Lei, W., Zhang, G., He, X., Miao, Y., Wang, X., Chen, L., & Chua, T.-S. (2020). Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* (pp. 2073–2083).

Li, Y., Capra, R., & Zhang, Y. (2020). Everyday cross-session search: How and why do people search across multiple sessions? In *Chiir '20, Proceedings of the 2020 conference on human information interaction and retrieval* (pp. 163–172). New York, NY, USA: Association for Computing Machinery.

Liu, Q., Cheng, G., Gunaratna, K., & Qu, Y. (2021). Entity summarization: State of the art and future challenges. *Journal of Web Semantics*, Article 100647.

Liu, J., Sarkar, S., & Shah, C. (2020). Identifying and predicting the states of complex search tasks. In *Chiir '20, Proceedings of the 2020 conference on human information interaction and retrieval* (pp. 193–202). New York, NY, USA: Association for Computing Machinery.

Matejka, J., Grossman, T., & Fitzmaurice, G. (2012). Citeology: Visualizing paper genealogy. In *Chi '12 extended abstracts on human factors in computing systems* (pp. 181–190). New York, NY, USA: ACM.

Mitsui, M., Liu, J., Belkin, N. J., & Shah, C. (2017). Predicting information seeking intentions from search behaviors. In *Proceedings of the 40th international acm sigir conference on research and development in information retrieval* (pp. 1121–1124). New York, NY, USA: ACM.

Peltonen, J., Belorustceva, K., & Ruotsalo, T. (2017). Topic-relevance map: Visualization for improving search result comprehension. In *Iui '17*, *Proceedings of the 22nd international conference on conference on intelligent user interfaces* (pp. 611–622). New York, NY, USA: ACM.

Rahdari, B., Brusilovsky, P., & Babichenko, D. (2020). Personalizing information exploration with an open user model. In *Ht '20*, *Proceedings of the 31st acm conference on hypertext and social media* (pp. 167–176). New York, NY, USA: Association for Computing Machinery.

Raman, K., Bennett, P. N., & Collins-Thompson, K. (2013). Toward whole-session relevance: Exploring intrinsic diversity in web search. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 463–472). New York, NY, USA: ACM.

Raman, K., Bennett, P. N., & Collins-Thompson, K. (2014). Understanding intrinsic diversity in web search: Improving whole-session relevance. *ACM Transanctions on Information Systems*, *32*(4), 20:1–20:45.

Reinanda, R., Meij, E., & de Rijke, M. (2015). Mining, ranking and recommending entity aspects. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 263–272).

Ruotsalo, T., Jacucci, G., & Kaski, S. (2020). Interactive faceted query suggestion for exploratory search: Whole-session effectiveness and interaction engagement. *Journal of the Association for Information Science and Technology*, *71*(7), 742–756.

Ruotsalo, T., Jacucci, G., Myllymäki, P., & Kaski, S. (2014). Interactive intent modeling: Information discovery beyond search. *Communications of the ACM*, *58*(1), 86–92.

Ruotsalo, T., Peltonen, J., Eugster, M. J. A., Głowacka, D., Floréen, P., Myllymäki, P., Jacucci, G., & Kaski, S. (2018). Interactive intent modeling for exploratory search. *ACM Transanctions on Information Systems*, *36*(4), 44:1–44:46.

Sarkar, S., & Shah, C. (2021). An integrated model of task, information needs, sources and uncertainty to design task-aware search systems. In *Ictir '21*, *Proceedings of the 2021 acm sigir international conference on theory of information retrieval* (pp. 83–92). New York, NY, USA: Association for Computing Machinery.

Shah, C., & White, R. W. (2021). Task intelligence for search and recommendation. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *13*(3), 1–160.

Shi, C., Zhang, Z., Luo, P., Yu, P. S., Yue, Y., & Wu, B. (2015). Semantic path based personalized recommendation on weighted heterogeneous information networks. In *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 453–462).

Shokouhi, M. (2013). Learning to personalize query auto-completion. In *Proceedings of the 36th international acm sigir conference on research and development in information retrieval* (pp. 103–112). New York, NY, USA: ACM.

Song, Y., Zhang, L., & Giles, C. L. (2011). Automatic tag recommendation algorithms for social recommender systems. *ACM Transactions on the Web (TWEB)*, *5*(1), 1–31.

Sordoni, A., Bengio, Y., Vahabi, H., Lioma, C., Grue Simonsen, J., & Nie, J.-Y. (2015). A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Cikm '15*, *Proceedings of the 24th acm international on conference on information and knowledge management* (pp. 553–562). New York, NY, USA: ACM.

Sun, J., Zhu, M., Jiang, Y., Liu, Y., & Wu, L. (2021). Hierarchical attention model for personalized tag recommendation. *Journal of the Association for Information Science and Technology*, *72*(2), 173–189.

Tang, L., Jiang, Y., Li, L., Zeng, C., & Li, T. (2015). Personalized recommendation via parameter-free contextual bandits. In *Proceedings of the 38th international acm sigir conference on research and development in information retrieval* (pp. 323–332). New York, NY, USA: ACM.

Teevan, J., Alvarado, C., Ackerman, M. S., & Karger, D. R. (2004). The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 415–422). New York, NY, USA: ACM.

Ukkonen, A., Joona, P., & Ruotsalo, T. (2020). Generating images instead of retrieving them: Relevance feedback on generative adversarial networks. In *Sigir '20*, *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval* (pp. 1329–1338). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3397271.3401129.

Vakkari, P., & Huuskonen, S. (2012). Search effort degrades search output but improves task outcome. *Journal of the Association for Information Science and Technology*, *63*(4), 657–670.

Verma, M., Yilmaz, E., Mehrotra, R., Kanoulas, E., Carterette, B., Craswell, N., & Bailey, P. (2016). Overview of the TREC tasks track 2016. In *Trec*.

Vuong, T., Andolina, S., Jacucci, G., & Ruotsalo, T. (2021). Does more context help? Effects of context window and application source on retrieval performance. *ACM Transactions on Information Systems*, *40*(2), http://dx.doi.org/10.1145/3474055.

White, R. W., Marchionini, G., & Muresan, G. (2008). Evaluating exploratory search systems. *Information Processing and Management*, *44*(2), 433.

Wildemuth, B. M., & Freund, L. (2012). Assigning search tasks designed to elicit exploratory search behaviors. In *Hcir '12*, *Proceedings of the symposium on human-computer interaction and information retrieval*. New York, NY, USA: Association for Computing Machinery.

Yee, K.-P., Swearingen, K., Li, K., & Hearst, M. (2003). Faceted metadata for image search and browsing. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 401–408). New York, NY, USA: ACM.

Yu, X., Ren, X., Sun, Y., Gu, Q., Sturt, B., Khandelwal, U., Norick, B., & Han, J. (2014). Personalized entity recommendation: A heterogeneous information network approach. In *Wsdm '14*, *Proceedings of the 7th acm international conference on web search and data mining* (pp. 283–292). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2556195.2556259.

Zamani, H., Dumais, S., Craswell, N., Bennett, P., & Lueck, G. (2020). Generating clarifying questions for information retrieval. In *Proceedings of the web conference 2020* (pp. 418–428). New York, NY, USA: Association for Computing Machinery.

Zamani, H., Mitra, B., Chen, E., Lueck, G., Diaz, F., Bennett, P. N., Craswell, N., & Dumais, S. T. (2020). Analyzing and learning from user interactions for search clarification. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval* (pp. 1181–1190). New York, NY, USA: Association for Computing Machinery.

Zhang, Y., Chen, X., Ai, Q., Yang, L., & Croft, W. B. (2018). Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th acm international on conference on information and knowledge management* (pp. 177–186). New York, NY, USA: Association for Computing Machinery.

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, *52*(1), 1–38.

Zhou, S., Dai, X., Chen, H., Zhang, W., Ren, K., Tang, R., He, X., & Yu, Y. (2020). Interactive recommender system via knowledge graph-enhanced reinforcement learning. In *Proceedings of the 43rd international acm sigir conference on research and development in information retrieval* (pp. 179–188).

Zuo, Y., Zeng, J., Gong, M., & Jiao, L. (2016). Tag-aware recommender systems based on deep neural networks. *Neurocomputing*, *204*, 51–60.