A Roadmap to User-Controllable Social Exploratory Search

CECILIA DI SCIASCIO, Know-Center GmbH, Austria
PETER BRUSILOVSKY, University of Pittsburgh, USA
CHRISTOPH TRATTNER, University of Bergen, Norway
EDUARDO VEAS, Know-Center GmbH, Austria

Information-seeking tasks with learning or investigative purposes are usually referred to as exploratory search. Exploratory search unfolds as a dynamic process where the user, amidst navigation, trial-and-error and on-the-fly selections, gathers and organizes information (resources). A range of innovative interfaces with increased user control have been developed to support exploratory search process. In this work we present our attempt to increase the power of exploratory search interfaces by using ideas of social search, i.e., leveraging information left by past users of information systems. Social search technologies are highly popular nowadays, especially for improving ranking. However, current approaches to social ranking do not allow users to decide to what extent social information should be taken into account for result ranking. This paper presents an interface that integrates social search functionality into an exploratory search system in a user-controlled way that is consistent with the nature of exploratory search. The interface incorporates control features that allow the user to (i) express information needs by selecting keywords and (ii) to express preferences for incorporating social wisdom based on tag matching and user similarity. The interface promotes search transparency through color-coded stacked bars and rich tooltips. This work presents the full series of evaluations conducted to, first, assess the value of the social models in contexts independent to the user interface, in terms of objective and perceived accuracy. Then, in a study with the full-fledged system, we investigated system accuracy and subjective aspects with a structural model that revealed that, when users actively interacted with all its control features, the hybrid system outperformed a baseline content-based-only tool and users were more satisfied.

CCS Concepts: • **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability;

Additional Key Words and Phrases: exploratory search, social search, user-controllable interface

ACM Reference Format:
Cecilia di Sciascio, Peter Brusilovsky, Christoph Trattner, and Eduardo Veas. 2018. A Roadmap to User-Controllable Social Exploratory Search. *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 1 (January 2018), 37 pages. https://doi.org/10.1145/3241382

1 INTRODUCTION

Learning and investigative tasks entail a combination of querying and browsing actions occurring as part of a dynamic process often called exploratory search [43]. Acquiring knowledge about a new topic is rarely fulfilled with a single query. Conversely, each bit of new knowledge triggers

Authors' addresses: Cecilia di Sciascio, Know-Center GmbH, Inffeldgasse 13, Graz, 8010, Austria, cdisciascio@know-center.at; Peter Brusilovsky, University of Pittsburgh, 135 N Bellefield Ave, Pittsburgh, 15213, USA, peterb@pitt.edu; Christoph Trattner, University of Bergen, Fosswinckelsgt. 6, Bergen, 5007, Norway, christoph.trattner@uib.no; Eduardo Veas, Know-Center GmbH, Inffeldgasse 13, Graz, 8010, Austria, eveas@know-center.at.

2160-6455/2018/1-ART1 $15.00
https://doi.org/10.1145/3241382

changes in information needs. In other words, learning is the result of a discovery cycle of several queries intermingled with the analysis of retrieved resources, where facts from a large volume of information fall in a conceptual representation [52].

It has been long acknowledged that exploratory search process can benefit from the "collective wisdom" of many people who can explicitly or implicitly collaborate in the search process [51]. Yet, almost all work on collaborative exploratory search focuses on explicit synchronous collaboration context [51, 64], with almost no studies of exploratory search interfaces based on "implicit collaboration" where future users can leverage information left by past users of information systems. This forms a sharp contrast with general stream of research on information retrieval where "implicit collaboration" known also as social search [4, 14, 27] has emerged into a broad stream of research. In this paper we attempt to bridge this gap by investigating the value of two social search approaches (one based on the use of social tags and another on collaborative user matching) in the context of exploratory search.

Our approach to support social exploratory search is to allow the user to filter documents by controlling the influence of search terms, as well to fuse traditional query-based document relevance with relevance produced by two "social" sources. To visualize this multi-source relevance, our interface uses visual cues to augment the ranked list, whereby relevance is not represented as a single score, but instead as the contribution of multiple dimensions, i.e. search terms and social sources. Intuitively, adding on control and transparency at multiple levels could possibly turn out into an overly complex system. Hence, our interest is to unveil whether interacting with a social exploratory search system based on multi-source relevance, like the one here proposed, can yield measurable benefits for users conducting exploratory search, in contrast to other simpler (but still interactive) systems.

Overall, the contributions can be summarized as follows:

(1) A hybrid content plus social ranking algorithm that accounts for fast changes in user needs and produces decomposable scores to facilitate representations in the UI.
(2) A comprehensive series of studies conducted at design stages of (2) and then with the fully integrated system described in (1). The validation roadmap of our approach to social exploratory search unfolded in three stages:
   (a) starting with assessment of classification and ranking accuracy using a static dataset,
   (b) followed by a crowd-sourced evaluation of accuracy as perceived by users rating 5-item ranked lists, and
   (c) finally evaluating the full-fledged system with users performing a more realistic exploratory search task.

This paper extends our previous work [20] with two evaluations that investigate the usefulness of the social models built in (2), which are learned from tagging data implicitly generated from bookmarking behavior of past users. By testing in experimental setups that are independent from the user interface, we seek to assess the individual performance of the social models in contexts that are unbiased from design implications of the UI. Hence, the two evaluations added in this version address stages (3a) and (3b). After introducing the social exploratory search system in section 3, sections 4 and 5 report on an offline experiment and a crowd-sourced evaluation, respectively, where we investigate the potential value of the proposed social models for exploratory search tasks. These two validation steps were conducted prior to the integration of social models in the search system. As it will be explained in section 6, they provide empirical justification for the conception of the hybrid user-controllable system described in section 3. The pay-off of these two evaluations at design stages is demonstrated in a final user study, reported in Section 7, revealing that users actively in control of the system obtain more accurate results, are able to appreciate transparency features and have a more satisfying searching experience.

2 RELATED WORK

2.1 Exploratory Search

Information seeking is a widely studied phenomenon [31, 57], as finding and organizing pieces of information occupies a large portion of our daily productive time. Information retrieval (IR) systems have grown as the preferred solution for contextualized search due to their ability to narrow down the number of entries to be inspected at a time. However, this kind of system requires precise user-generated queries. As the user learns about certain topic, queries have to be iteratively reformulated to express evolving information needs. Formulating queries has proven to be more complicated for humans than plainly recognizing information in a visual manner [31], which is why the combination of IR with HCI techniques has led towards a shift in the way users search. Browsing search strategies, which rely on on-the-fly selections, navigation and trial-and-error, are associated with the term that Marchionini et al. [43] coined "exploratory search". By definition, exploratory search is open-ended, i.e. the user starts the search with some initial query in mind and discovers the next query terms along the way, as she finds and scrapes new information.

Over the last decades, several approaches attempted to foster deeper exploration and sensemaking of search results or large document collections. Ranked lists have prevailed as the paradigm for presenting results, due to their familiar format and because users know where to start inspecting items. Although ranked lists alone are regarded as opaque and under-informative [30], Shani et al. [56] suggest that: (i) explaining relevance scores encourages users to explore beyond the first two results, and (ii) users prefer bars over numbers or the absence of graphical explanations. In addition, the use of visualizations has been favored due to their capacity to convey document relevance by exploiting pre-attentive patterns. For example, tile bars [30] encode relative frequency of query terms with compact shaded blocks, whereas other approaches complement lists with visual metaphors [47], similarity-preserving layouts [28] or POI-based visualizations [48].

Unlike general-purpose search systems, e.g. Google, the trend in dedicated exploratory search systems (ESS) is headed towards highly interactive user interfaces (UIs) supported by ever-growing artificial intelligence methods. Examples of such features include task models [3], categorized overviews [61], adaptive visualizations [1] and interactive intent modeling [54]. Similar to other research approaches, our work extends the popular faceted search [63] to a context where filtering methods function independently from metadata-based information. In this group of UIs, the original query is used to produce not only a list of results, but also a list of most important information entities covered by the retrieved results – from simple unigrams [21, 33] to named entities [2] and user-specified keywords [54]. These entities uncover the *aboutness* of the results and serve as interactors for further exploration of the generated ranked list.

More specifically about the *uRank* interactive ranking approach [21], we replicate features which allow the user to manipulate the weights of the selected entities to adjust the original ranking to their emerging needs. This system extracts keywords from titles and abstracts and displays them in the UI as interactors. As the user selects (or types) keywords of interest, a document ranking visualization is re-sorted in real-time, thus promoting a *search-by-browsing* information access paradigm. di Sciascio et al. found that search-by-browsing supported by this kind of adaptive system incurs in lower cognitive load without hindering user performance, compared to conventional list-based UIs [22].

After several users have conducted exploratory search with a system, they leave traces behind that provide hints as to what combinations of terms were fruitful in their searches, or which documents were preferred by users with similar information needs. This kind of traces are the means to support future users, by complementing exploratory search with social wisdom.

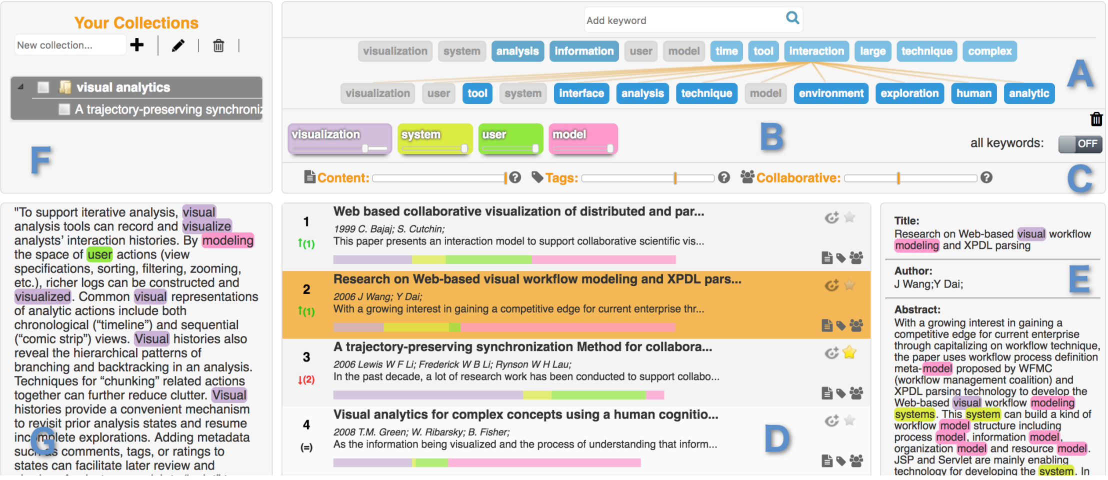
2.2 Social Search and Interactive Fusion of Relevance Sources

Social search is the common name for a group of information retrieval approaches that use traces left by past users of information systems to help future users in the search process [14, 24, 27]. The traces of past users can be obtained from multiple sources, such as search logs, social tagging systems, Web site logs, and various social media systems. In turn, this information could be leveraged to assist users in different stages of their search including query formulation [9], matching and ranking of results [10], and augmentation of results [4]. The most popular target of social search techniques is ranking of search results. The idea of social ranking is to combine traditional content-based relevance of search results with relevance measures obtained from social sources. For example, documents that are frequently selected in response to similar queries as well as results that have been recognized by target users by bookmarking, tweeting, or other kinds of sharing in social systems have high social relevance and have to be promoted in the list of search results. Traditionally, the fusion of content-based and social relevance is done automatically by learning the weights of different sources using learning-to-rank or similar data-driven approaches [37]. The automatic fusion, however, does not fit well to the nature of exploratory search where the user might want to decide to what extent social information should be considered to rank results of a particular query. In our paper, we suggest an interactive controllable fusion of content-based and social relevance judgment and present an interface that allows users to control this fusion in a way that is consistent with modern information exploration interfaces.

Interactive fusion of relevance sources is not an entirely new topic in the field of information access and intelligent interfaces. Most research on this topic, however, has been done in the area of hybrid recommender systems (RS) [15]. In the context of so-called parallel hybridization, a RS has to fuse relevance judgment obtained from difference sources or approaches. Just like in the case of social search, the traditional approach to source fusion in hybrid systems is automatic, where the influence of each component is determined using some machine learning approach and stays the same. This, however, does not allow for accommodating to real-life context where the importance of each source could depend on the varying user needs. For example, a movie recommendation could be a hybrid of personal collaborative filtering and social recommendation collected through social connections. When watching alone, a user may put more emphasis on the personal part, while when selecting a movie to watch with friends, the social part should be more valuable.

User-controlled source fusion is part of a broader stream of work on controllability for RS. A range of user studies on interactive and controllable RS indicates that when users are given the chance, they do make use of their ability to control the system [23] and are more satisfied [11]. For example, *TasteWeights* [11] generates music recommendations from multiple sources and presents them in a visually rich interface that allows the user to interactively change the contribution of each individual source to the ranked list. *SetFusion* [50] allows to control the contribution of three sources in a hybrid RS for research talks and indicates which sources were used to recommend each item with color-coding. For further details, He et al. [29] elaborated a comprehensive survey of interactive RS, including visualization, presentation and interaction aspects.

Within the area of information retrieval (IR), the idea of interactive fusion attracted much less attention, since the presence of several independent relevance sources is less typical. Also, extensive exploration with an interactive source-fusion interface is not natural in most search contexts. One exception is the work on personalized search, whereby each search result has two relevance scores - relevance to the query and to the user profile. Traditionally, personalized search systems perform automatic fusion of query-based and profile-based relevance rating to offer personalized ranking



"To support iterative analysis, visual analysis tools can record and visualize analysts' interaction histories. By modeling the space of user actions (view specifications, sorting, filtering, zooming, etc.), richer logs can be constructed and visualized. Common visual representations of analytic actions include both chronological ("timeline") and sequential ("comic strip") views. Visual histories also reveal the hierarchical patterns of branching and backtracking in an analysis. Techniques for "chunking" related actions together can further reduce clutter. Visual histories provide a convenient mechanism to revisit prior analysis states and resume incomplete explorations. Adding metadata such as comments, tags, or ratings to states can facilitate later review and

1 ↑(1) **Web based collaborative visualization of distributed and par...**
1999 C. Bajaj; S. Cutchin;
This paper presents an interaction model to support collaborative scientific vis...

2 ↑(1) **Research on Web-based visual workflow modeling and XPDL pars...**
2006 J Wang; Y Dai;
With a growing interest in gaining a competitive edge for current enterprise thr...

3 ↓(2) **A trajectory-preserving synchronization Method for collabora...**
2006 Lewis W F Li; Frederick W B Li; Rynson W H Lau;
In the past decade, a lot of research work has been conducted to support collabo...

4 (=) **Visual analytics for complex concepts using a human cognitio...**
2008 T.M. Green; W. Ribarsky; B. Fisher;
As the information being visualized and the process of understanding that inform...

Title:
Research on Web-based visual workflow modeling and XPDL parsing

Author:
J Wang;Y Dai;

Abstract:
With a growing interest in gaining a competitive edge for current enterprise through capitalizing on workflow technique, the paper uses workflow process definition meta-model proposed by WFMC (workflow management coalition) and XPDL parsing technology to develop the Web-based visual workflow modeling systems. This system can build a kind of workflow model structure including process model, information model, organization model and resource model. JSP and Servlet are mainly enabling technology for developing the system. In

Fig. 1. The UI of the social exploratory search system provides features to control a hybrid ranking model and graphic explanations of the information source provenance.

[45]. However, Ahn et al. [3] demonstrated that allowing users to choose how to combine these two ratings across individual queries can increase all performance aspects.

Social exploratory search offers an excellent opportunity to apply interactive fusion of relevance sources outside of the typical recommendation context. On one hand, the social search approach provides several relevance sources beyond regular query-based relevance (each type of socially collected information could be used for independent relevance ranking). On the other hand, the context of exploratory search makes interactive exploration of results with user involvement into ranking quite natural.

In light of the little work done on controllable fusion of multiple sources of relevance in IR and the interactive nature of the exploratory search task, we propose a rich user interface that allows the user to combine traditional query-based relevance with two "social" relevance sources - one based on user-approved tags, and one based on collaborative matching between the current and past users. The next section describes control and transparency features implemented in the UI and the ranking model supporting the exploratory search task. The following studies investigate the potential value of using social relevance models for this kind of task, in experimental conditions that are independent and unbiased from design aspects of the UI. Thereafter, we disclose the benefits on system performance and user experience when users search with the full-fledged system.

3 HYBRID URANK: A USER-CONTROLLABLE SOCIAL EXPLORATORY SEARCH SYSTEM

In order to blend exploratory and social search, we replicate features of *uRank* [21], an adaptive system designed for exploratory search (ES) of textual documents. The basic system promotes a search-by-browsing information access paradigm, using keywords extracted from search results as interactors to refine a document ranking upon evolving information needs.

As the user interacts with extracted keywords and bookmarks resources, our system can learn about the importance of certain keywords and their connection to documents. As a result, the social-enhanced ES system incorporates collaborative and tag-based filtering methods into a hybrid ranking model, whereby (i) tagging data used for training purposes is implicitly generated from bookmarking behavior, and (ii) the interface allows the user to tune the fusion between content

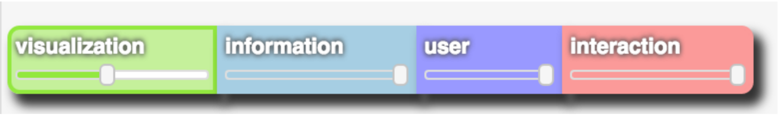
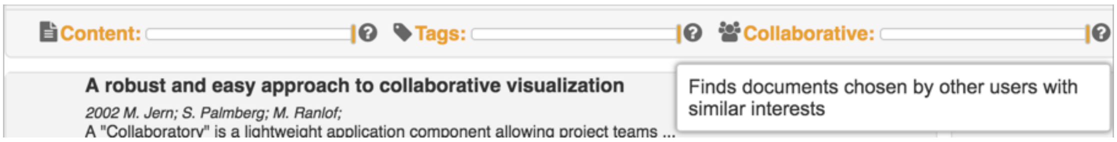


Fig. 2. Switching on the "All Keywords" filter sets a strict "and" constraint. Keyword tags in the *Query Box* appear grouped in one body (replacing individual elements in Figure 1.B).



A robust and easy approach to collaborative visualization

2002 M. Jern; S. Palmberg; M. Ranlof;

A "Collaboratory" is a lightweight application component allowing project teams ...

Finds documents chosen by other users with similar interests

Fig. 3. Fusion of Relevance Sources. The UI incorporates sliders to control the relative weight of each ranking method. Informative tooltips appear by hovering over question mark icons.

information and *social wisdom* in the resulting ranking. In the remainder of this section we briefly describe the features of the user interface (Section 3.1) and the elements of the system that enable social search functionalities alongside exploratory search.

3.1 The User Interface

Exploratory search is often motivated by a complex information problem where the user has little understanding of terminology or information space structure [60]. Our tool supports the information seeking process with features for exploration and explanation, which largely correspond to control and transparency in the UI. The motivation for control and transparency features comes from research on user experience with recommender systems, whereby users can gain in confidence and performance when the system does not behave like a black box but is instead flexible to users' preferences [11] and provides sufficient explanations [32].

As a starting point for exploration, the 12 most frequent keywords appear arranged in an inline fashion, at the top of the UI in the *Keyword Box* (Figure 1.A). By hovering on a first-row keyword, the user can discover other keywords frequently appearing together in the documents' text, which appear on the second row (connected by tree-shaped links).

3.1.1 Controlling the System. The *Query Box* (Figure 1.B) is the component that captures the user's current information needs, expressed in terms of iterative interactions with keywords. Possible interactions inherited from the original system include keyword addition (either by selection ot manually typing), deletion and weighting. We extended these options with filtering functionalities, enabling users to: (i) filter individual keywords, so that bearing documents remain in focus and the rest are dimmed (▼ button visible on mouse over); and (ii) enable the "all keywords" filter (toggle button on the right side of the *Query Box*). The latter switches to a strict bearing criterion (from "or" to "and"), so that the ranking includes only documents where all selected keywords appear in the text. Selected terms in the *Query Box* are rendered as a joint, single body, as shown in Figure 2.

The main addition to the UI in order to support social features is the ranking control area, illustrated in Figure 3 (panel C in Figure 1). Ranking controls have a double function: informative and control-enabling. On one hand, they intend to make the user aware of the criteria applied by the system to rank documents. On the other hand, and more importantly, these controls allow the user to modify the impact of content-based and social relevance in the ranking. Thus, ranking sliders adjust the weights of the three methods in the hybrid model:

- **Content**: ranks documents based on terms contained in their titles and abstracts.
- **Tags**: based on selected keywords matching tags associated to documents by bookmarking actions of past users.
- **Collaborative**: brings documents previously bookmarked by other users with similar interests.

3.1.2 Transparency in the UI. A highly controllable UI could be cumbersome if users cannot perceive the effect of their actions. Therefore, we rely on graphical explanations to convey system decisions and mitigate the complexity of multiple control features. Taking into account that the tool allows for tuning both keywords and model parameters, we had limited options for the use of color to convey score contributions. We decided to use a categorical color palette to represent keywords and thus maximize the amount of information encoded with color. Color-coded keywords are visual cues that pop out from their surroundings, enabling the user to pre-attentively recognize them in the text and perceive their general context prior to conscious reading.

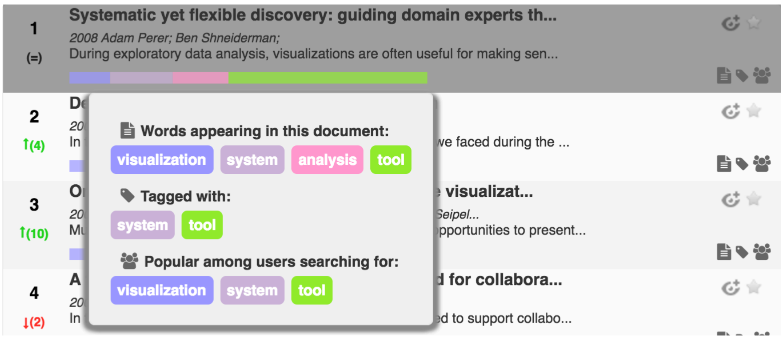
Color-coded stacked bars embedded in the *Document Ranking* (Figure 1.D) indicate document scores. The overall width is the total score, while single bars in the stack denote individual keyword contributions. This means that the score produced by each method in the ranking model is broken down into its constituent keyword-based sub-scores. Sub-scores are then added keyword-wise to represent a single bar in the stack. For example, in Figure 4 the length of the green bar is obtained as the addition of the content, tag and collaborative sub-scores for the keyword "tool".

In turn, icon hints at the bottom-right corner of individual items in the *Document Ranking* reveal at a glance which methods ranked a given document. The icons match the labels in the model parameter controls (Figure 1.C). Visualizing the overlap between keywords and information source is possible by hovering over the stacked bars. The tooltip in Figure 4 explains which keywords appear in the document, which ones have been used to tag the document and which ones are popular among users with similar interests.

When the user clicks on a list item, the *Document Viewer* (Figure 1.E) displays the available text and metadata information for the selected document. Terms matching selected keywords in the *Query Box* are highlighted in the same colors, enabling the user to readily spot them in the text prior to conscious reading. The same principle is applied for the notepad in Figure 1.G. As the user saves a document into a collection (organized in the *Collection Panel*, in Figure 1.F), this piece of text is augmented with colored terms, which are related to bookmarked resources.

3.2 Generating Tagging Data

Text-mining or topic-modeling methods generate machine-based content descriptions, often referred to as keywords or key phrases (in the case of unigrams and n-grams, respectively). Conversely,


1 (=) Systematic yet flexible discovery: guiding domain experts th...
2008 Adam Perer; Ben Shneiderman;
During exploratory data analysis, visualizations are often useful for making sen...

Words appearing in this document:
visualization system analysis tool
Tagged with:
system tool
Popular among users searching for:
visualization system tool

Fig. 4. Hovering over a document's stacked bars shows a tooltip informing which tags influenced each RS.

tags are regarded as human-generated descriptors given by one or more users. Social tagging data is the space where relationships between users, resources and tags are captured. Unlike a typical social tagging system, which serves as a platform for users to tag resources, our system is designed for users to search and explore documents. Hence, our approach to incorporate social information does not work with social tagging data in the strict sense, as users do not explicitly assign tags to documents at any time. Instead, we leverage bookmarking behavior to infer human-approved content descriptors, which we treat as (so-called) tags, i.e. tags are not generated but rather approved by humans.

As the user explores search results by selecting extracted keywords from the *Keyword Box*, they save relevant documents along the way as bookmarks. Bookmarking actions provide useful information in a twofold manner: *i)* the user establishes an explicit preference for a given document, and *ii)* the user implicitly considers that the document is relevant for their current information needs, which are expressed by the selected terms present in the *Query Box* at the moment of the bookmark event. We can assume that a user regards the currently selected terms as good content descriptors for the bookmarked document, and hence refer to them as tags in our tagging space. One reason to treat human-approved content descriptors as tags is that humans largely tend to generate keyphrases that already appear in the text [17]. For clarification, in the remainder of this paper we refer to keywords or terms as automatically extracted unigrams, and recall tags as terms used to implicitly tag bookmarked documents. Therefore, a document is tagged with *tags* $t_1, \cdots, t_n$ if a user bookmarks the document after selecting a set of *terms* $t_1, \cdots, t_n$ in the UI.

Formally, B is the set of all bookmarks, where a single bookmark is represented by a triple of the form $(u, d, Q) \in B$. Thus, a bookmark entry denotes that user u saved a document d and implicitly tagged it with selected terms $Q = \{t_1, \cdots, t_n\}$. Hereby, we define the folksonomy $\mathcal{F} = \langle U, D, T, A \rangle$, where U is the set of all users, D the set of all documents, T the set of all tags and A (Formula 1) is a function mapping a bookmark $b \in B$ into a set of $|Q|$ triples of the three entity types, (u, d, t).

$$A = \{ \ (u, d, t) \mid u \in U, \ d \in D, \ t \in Q \subseteq T, \ \exists \ (u, d, Q) \in B \ \} \tag{1}$$

Triples mapped by A conform the tagging space that represents ternary relationships between users, documents and tags. We break down ternary relationships into the following 3 matrices:

- $R \in [0, 1]^{|U| \times |D|}$ is similar to the user-item matrix employed in CF. Unlike rating data, where values typically range from 1 to 5, in this context user preference for an item is binary, thus $r_{ud} = 1$ if user u bookmarked document d, 0 otherwise.
- $X \in \mathbb{N}^{|D| \times |T|}$ is the document-tag matrix, such that the i^{th} row represents the profile vector for the i^{th} document, in terms of tags associated to it at the moment it was bookmarked.
- $Y \in \mathbb{N}^{|U| \times |T|}$ is the user-tag matrix, such that y_{ut} indicates the number of documents bookmarked by user u after selecting term t (and implicitly using t as a tag).

In a nutshell, we extract explicit and implicit information contained in bookmarking behavior to generate tagging data, which is then employed to train the social-based ranking methods described in the next section.

3.3 Ranking Model

As user u adds a set of terms $Q = \{t_1, \cdots t_n\}$ to the *Query Box*, the system computes independent document scores with three different models and ranks them according to an overall hybrid score. The content-based model produces scores based on textual information from documents' titles and abstracts. This is the ranking method in the original system. The tag-based and collaborative models are incorporated to extend the original system and support social exploratory search. In this section we describe the three models and the hybridization of the final score. Interaction and presentation factors steered the design and implementation of these methods, namely:

(1) The models have to adapt to sudden changes in information needs by computing scores on-demand. That is to say, they should provide a sense of real-time responsiveness upon user interactions with keywords.
(2) Then for presentation purposes, they should generate decomposable scores that can be graphically explained, i.e. calculate separate scores for each selected term $t \in Q$.

3.3.1 Content-based (CB). CB ranks documents by the relative frequency at which the selected tags (keywords) appear in titles and abstracts[1]. We build a vector space model and compute document-query similarity as the weighted cosine similarity (Formula 2).

$$s_{cb}(d,Q) = \frac{1}{|Q|} \sum_{t \in Q} \frac{tfidf(t,d) \ \cdot \ w_{ut} \cdot \Gamma(d,Q)}{\parallel d \parallel \cdot \parallel Q \parallel}, \tag{2}$$

where $tfidf(t,d)$ is the tf-idf [59] score for term t in document d, w_{ut} is the weight assigned to term t by user u in the UI, while $||d||$ and $||Q||$ are the Euclidean norms for vectors d and Q, respectively. Lastly, $\Gamma(d,Q) = exp(\alpha\ (|d \cap Q| - |Q|))$ is a decay function that penalizes documents not containing all selected keywords. We set the α parameter to 0.25 to soften the decay rate.

3.3.2 Tag-based (TB). This model measures the strength between the target user's selected terms Q and a document d in the social tagging space. In other words, the model expresses to what extent terms Q are good content descriptors of d from the perspective of other users that bookmarked (and implicitly tagged) d in the past.

Given the document-tag matrix $X := (x_{td})$, the similarity between term t and document d, $sim(t,d)$ is computed as the conditional probability of d to be bookmarked given t, which in Formula 3 is represented as the ratio between the tagging frequency of d under t and the tagging frequency for t across all documents. s_{tb} is then the probability of t and d occurring together, i.e. $p(t) \cdot p(d \mid t) = p(t \cap d)$.

$$s_{tb}(u,d,Q) = \sum_{t \in Q} w_{ut} \cdot \frac{x_{td}}{\sum\limits_{d' \ \in \ D} x_{td'}} \tag{3}$$

3.3.3 User-based (UB). The nature of this model resembles that of a collaborative filtering recommender, except that traditional CF was conceived to learn a model from rating data, whereas ratings in the case of bookmarks is binary. i.e. a document is either bookmarked or not. The UB model estimates the likelihood for a document d to be bookmarked by target user u based on the strength between u's neighborhood V and d. Since u's tag preferences are frequently updated through the UI, similarity between two users is in this case agnostic of u's past search interests. Instead, v is similar to u if v has bookmarked any document with selected terms Q in the past.

Given the user-tag matrix $Y := (y_{ut})$, we calculate the similarity between u and v as the weighted ratio between the times v used t and the total times t was used by any user to bookmark any document, as denoted in Formula 4.

$$sim(u,v) = \sum_{t \in Q} w_{ut} \cdot \frac{y_{vt}}{\sum\limits_{v' \in V} y_{v't}} \tag{4}$$

[1] Ranking by full text is also supported. But, since keyword extraction results tend to be much noisier with full text, we only indexed titles and abstracts for the subsequent studies.

The neighborhood for u, V is formed as the union set of the neighborhood for each $t \in Q$, V_t (Formula 5), such that, V_t is the set of users that implicitly tagged any document with term t. V is then trimmed to the k top neighbors.

$$V = \bigcup_{t \in Q} V_t \mid v \in V_t \; if \; y_{tv}^{\intercal} > 0 \tag{5}$$

The user-item matrix R is extended into $\tilde{R} \in \mathbb{R}^{|U| \times |D|}$, such that the preference of user v for document d is 1 if v bookmarked d, $sim_J(v, d)$ otherwise, where $sim_J(v, d)$ is the Jaccard coefficient between v's user profile and the d's document vector. The final s_{ub} score is obtained as shown in Formula 6 by averaging the product of user-neighbor and neighbor-document similarities across all neighbors in V.

$$s_{ub}(u, Q, d) = \frac{1}{|V|} \sum_{v \in V} sim(u, v) \cdot r_{vd} \tag{6}$$

3.3.4 Hybrid Overall Score. Relative ranking weights can be interactively adjusted in the $[0, 1]$ range and are represented by vector ω. ω is balanced with the softmax function, so that $\sum_{i \in |\omega|} \omega_i = 1$. The outputs of the three models are first min-max normalized to avoid that higher scores from one method undermine the contribution of the others. Then, the hybrid score for d is the linear combination shown in Formula 7, where $\hat{s}_{cb}$, $\hat{s}_{tb}$ and $\hat{s}_{ub}$ are the normalized values for the corresponding scores.

$$s(u, Q, d) = \omega_{cb} \; \hat{s}_{cb}(d, Q) + \omega_{tb} \; \hat{s}_{tb}(u, d, Q) + \omega_{ub} \; \hat{s}_{ub}(u, d, Q) \tag{7}$$

To allow for explanations of individual keyword contributions, all methods calculate separate scores for each $t \in Q$, which are then fused and represented as color-coded bars in the UI.

The three methods in the hybrid model were chosen as a result of the evaluations described in the next two sections (4 and 5), whereby we assessed objective and perceived accuracy of the TB and UB models, in a step that preceded their integration in the hybrid user interface. We contrast the findings for the two experimental setups in Section 6, sustaining our decision of employing both, the TB and UB models, in our approach to social exploratory search. The full-fledged system herein described is consequently evaluated in Section 7.

4 OFFLINE EXPERIMENT: OBJECTIVE ACCURACY

This experiment takes the first step to assess how social search methods based on implicit collaboration learned from bookmarking behavior perform in comparison to the CB originally implemented in *uRank* and a baseline most-popular model. In order to account for contextual aspects, we used a dataset comprising bookmarks generated during a previous user study [22]. This dataset incorporates information about the type of exploratory search task the subjects were conducting, namely focused or broad.

For this experiment, we also included a merged version of the TB and UB models, namely TU_β (Section 4.1), whereby β is a parameter that controls the amount of tag and user information in the hybrid score. The results in this section first report on intra-contrasts, in order to determine the right blend of the tag-based and user-based models in the combined model (TU_β). In other words the purpose was to find the optimal value for the β parameter. Next, item-prediction is addressed by evaluating classification accuracy of the three social models and the two baselines. Finally, the influence of the type of search task is analyzed.

4.1 Tag+User Model (TU$_\beta$)

Given a target user u and a set of preferred keyword tags Q, TU_β computes an overall document score, s_{tu}, as the weighted sum between outputs from the *TB* and *UB* models, as illustrated in Formula 8. β is a parameter for adjusting the level of hybridization between both score types. The second line of the Formula expresses s_{tu} in terms of probabilities, emphasizing the incorporation of explicit user preference in the form of weighted query terms.

$$s_{tu}(u,Q,d) = \sum_{t \in Q} \beta \cdot s_{tb}(u,Q,d) \; + \; (1-\beta) \cdot s_{ub}(u,Q,d) =$$
$$\sum_{t \in Q} \underbrace{p(t)}_{\text{u's pref.}} \cdot \left[\beta \; \underbrace{p\,(d \,|\, t)}_{\hat{s}_{tb}} + (1-\beta) \; \underbrace{\frac{1}{|V|} \sum_{v \in V} p(v \,|\, t)\; \tilde{r}}_{\hat{s}_{ub}} \right] \tag{8}$$

Although the algorithm produces values between 0 and 1, in practice the tag-based model tends to produce higher scores that undermine the contribution of user-based values, even in a balanced configuration ($\beta = 0.5$). Therefore, the underbraced $\hat{s}_{tb}$ and $\hat{s}_{ub}$ sub-scores in Formula 8 are normalized across all documents in D in advance.

4.2 Experimental Setup

This evaluation was conducted to investigate the performance of social relevance models learned from implicit user collaboration. The assessment takes a content and a popularity-based model as baseline, whereby comparisons are drawn in line with focused and broad exploratory search tasks.

Validation was conducted by feeding the five models with batches of training and test sets (B_T and B_V, respectively) and computing classification accuracy $@k$, with k ranging from 1 to 5. A single batch consisted of a 70-30 fold sample, which means 994 bookmarks were assigned to training and the remaining 426 items to testing. The procedure was repeated in a 10-fold cross-validation on hold-out sets. In order to generate the training set B_T , we split bookmark logs (each comprising one user, one document and all selected tags) into triples of the form $\mathcal{A} = (u, d, t)$. These triples were then used to populate matrices R, X and Y.

Prior to analyzing differences across the five models, we first investigated different configurations of the combined social model TU_β, in order to find the optimal blend between user and tag information. We fixed the neighborhood size fixed to 10, cross-validation on hold-out sets was executed, whereby the β parameter was adjusted from 0.1 to 0.9, with an incremental step of 0.1. Thus, the tested variations ranged from almost pure user-based ($\beta = 0.1$) to almost pure tag-based ($\beta = 0.9$), with seven hybrid combinations of both score types in between ($\beta = .2, \cdots, .8$).

For the overall analysis, the three variations of social search were included, namely: pure tag-based (*TB*, $\beta = 1$), pure user-based (*UB*[2], $\beta = 0$) and a balanced hybrid model ($TU_{\beta=.5}$, $\beta = 0.5$). Thus, cross-validations on hold-out sets compare the social models against the two baseline approaches: content-based (*CB*) and most-popular (*MP*). Furthermore, the influence of the type of search task was analyzed by incrementally partitioning the test set B_V by *focused* and *broad* search.

4.2.1 *Dataset.*

The dataset was built from bookmarks collected during the evaluation reported in [22]. Tasks were developed under a simulated exploration scenario: the participant received a list of documents related to a given topic and the goal was to select the 5 most relevant documents for a given piece of text.

[2] Neighborhood size was set to 10, in order to compensate for the small pool of users in the experiment dataset.

The study was structured in a within subjects design, whereby participants performed four iterations of the same task with either *uRank* (U) or a baseline SERP-like UI (L for *List*). Furthermore, participants were presented with result lists of different sizes, i.e. either 30 or 60 items at a time. Therefore, the study followed a a 2 x 2 repeated measures design with *tool* and *#items* as independent variables, each with 2 levels (*tool* = U/L, *#items* = 30/60). Participants had to perform a total of four tasks, each under one of the possible combinations of the independent variables, such that the ordering of conditions was randomized with balanced Latin Square. To counterbalance learning effects, we prepared four datasets covering a variety of topics: *Robots*, *Augmented Reality*, *Women in Workforce* and *Circular Economy*. *Topic* was treated as a random variable within constraints and assigned at random across tasks.

Each single task comprised three sub-tasks: two focused exploration sub-tasks (Q1 and Q2) and a broad exploration sub-task (Q3). For Q1 and Q2, participants had to find relevant documents for two or three given keywords. In turn, Q3 was about finding five items relevant to a short text extracted from the *Wikipedia* page for the given topic. The focused tasks attempted to reflect the behavior of quickly shifting information interests within a topic while exploring, whereas Q3 required the user to clarify a textual description and deduce terms and phrases to shape information needs.

The study unfolded in a controlled environment, where 24 participants took part (11 female, 13 male, between 22 and 37 years old). We mostly recruited them from the medical and computer science student population at Technical University of Graz. We corroborated that none of them was knowledgeable in the topic areas selected for the study. A total of 1420 bookmarks were collected at the time. For this offline experiment, we generated bookmark tuples of the type (u, d, Q) in order to populate the ground-truth data set B_T.

An important remark is that words belonging to the same family were grouped under a common super tag, e.g "robots" and "robotics" are represented by "robot". The same applies to variations of British and American English, e.g. "colour" and "color". Stemming and selection of representative terms is explained for the keyword extraction module in [21], which produces the terms employed for interaction in the UI and posterior tagging of bookmarked resources.

Data Quality. We analyzed the "consensus" among participants as a proxy to assess the quality of the ground truth dataset. To do so, we aggregated the collections gathered by participants for all conditions, and then computed cosine similarity across *tool*, *#items*, topic (WW, Ro, AR, and CE) and sub-task (Q1, Q2 and Q3). Overall, participants' choices regarding relevant documents matched approximately three out of four times ($M = .73,\ SD = .1$). Table 1 breaks down similarity values for the two different *tools* (U vs. L) across sub-task and topic. On average, focused searches (Q1 and Q2) ($M = .77,\ SD = .13$) tended to reach higher consensus than broad searches (Q3) ($M = .66,\ SD = .13$). Similarity also remained high for bookmarks across the four topics, with

Table 1. Cosine similarity between bookmark collections gathered with *uRank* and *List* conditions in a previous study. Bookmarked items were used as ground truth in the current offline experiment.

Task Type	WW	Ro	AR	CE	All topics
Q1 (focused)	.55	.79	.58	.74	.66
Q2 (focused)	.70	.86	.84	.86	.81
Q3 (broad)	.75	.72	.75	.63	.72
All Tasks	.66	.79	.72	.74	.73

Note: WW = Women in Workforce; Ro = Robots; AR = Augmented Reality; CE = Circular Economy

Robots as the top and *Women in Workforce* as the least uniform. For a more detailed analysis of across topics and task type, refer to [22].

4.2.2 Performance Metrics. Performance was evaluated in terms of state-of-the-art accuracy metrics for information retrieval and recommender systems. Classification accuracy was measured with *recall*, *precision*, *F-measure*; while *normalized Discounted Cumulative Gain* (*nDCG*) and *Mean Reciprocal Rank* (*MRR*) were used for ranking accuracy.

The test set B_T employed in the study contains bookmark logs that represent a positive decision from a user with respect to a document. Therefore, no true-negatives are available and accuracy computation relies on "hit" count. A single test consisted in pairwise comparisons between a bookmark log in B_T and the list of top-k documents obtained for the given user and keyword tags. Hence, let $rank_d$ be the rank of bookmarked document d in the list, then a hit occurs when $rank_d \leq k$. The formulas for *recall*, *precision* and *F-measure* have been adjusted accordingly.

While classification accuracy considers hits irrespective from rank, ranking accuracy metrics penalize items that fall farther from the top position. Formulas for *nDCG* and *MRR* are thus adjusted to iterate across each entry in the validation set rather than across each user and document.

Recall@k. *Recall* is computed as the ratio between correctly retrieved items and all relevant items. As Cremonesi et al. [19] explain it, recall for a single test can assume either 0 (in case of miss) or 1 (in case of hit), because there is only one relevant item per test. Thus, overall recall@k (Formula 9) is computed by averaging the total number of hits over all cases in the test set B_V.

$$Recall@k = \frac{\#hits}{|B_V|} \tag{9}$$

Precision@k. Since this metric is the ratio of retrieved items that are relevant, *precision* for a single test can take either the value 0 (miss) or $1/k$ (hit). Then, *Precision@k* is the average of the single tests (Formula 10). Note that the assumption that, when a hit occurs, all $k-1$ non-selected items are irrelevant to user u tends to underestimate the computed measures with respect to the actual precision and recall.

$$Precision@k = \frac{\#hits}{|B_V| * k} = \frac{Recall@k}{k} \tag{10}$$

F1@k. Formula 11 corresponds to a balanced F measure:

$$F1@k = 2 * \frac{Precision@k * Recall@k}{Precision@k + Recall@k} \tag{11}$$

nDCG@k. The *Discounted Cumulative Gain* (*DCG*) is given by Formula 12, where $rank_d$ corresponds to the position for document d in the result list, and B is a function that return 1 if the document at position i_d is relevant. In any case, $B(rank_d) = 0 \; if \; rank_d > k$.

$$DCG@k = \frac{1}{|B_V|} \sum_{d \, \in \, B_T} \frac{2^{B(rank_d)} + 1}{\log_2 (rank_d + 1)} \tag{12}$$

Then, *Normalized Discounted Cumulative Gain* [34] (*nDCG@k*) is calculated as the ratio between *DCG@k* and the ideal *DCG* value, *iDCG@k*, which is the highest possible *DCG* value that can be achieved if all the relevant items are ranked in the correct order. Note that in the case of pair-wise comparisons against true-positive samples, at most one item in the list can be relevant, thus *iDCG* is always 1 and the above formula for *DCG@k* is inherently normalized. Herein, this section refers to Formula 12 as *nDCG@k*.

MRR@k. *Mean Reciprocal Rank* is a widely adopted metric in information retrieval and it applies to scenarios in which only one item in the list is relevant, as it is in the current case. This measure calculates the reciprocal of the rank at which the first relevant document was retrieved and averages the reciprocal ranks across all bookmarks in the test set. Assuming that the user will look down the ranked list until a relevant document is found, and that the document is at $rank_d$, then the precision of the set they view is $1/rank_d$, which is also the reciprocal rank measure. For this reason, *MRR* is equivalent to *Mean Average Precision* (*MAP*) in cases where each query has precisely one relevant document [18].

$$MRR@k = \frac{1}{|B_V|} \sum_{d \ \in \ B_T} \frac{1}{rank_d} \tag{13}$$

4.2.3 Baseline Algorithms. Since the social models are designed for exploratory search, one of the baselines in the experimental setup is the content-based method used in the original *uRank* tool. Additionally, we tested against a popularity-based model, as it is usually done in evaluations of collaborative or social recommender systems.

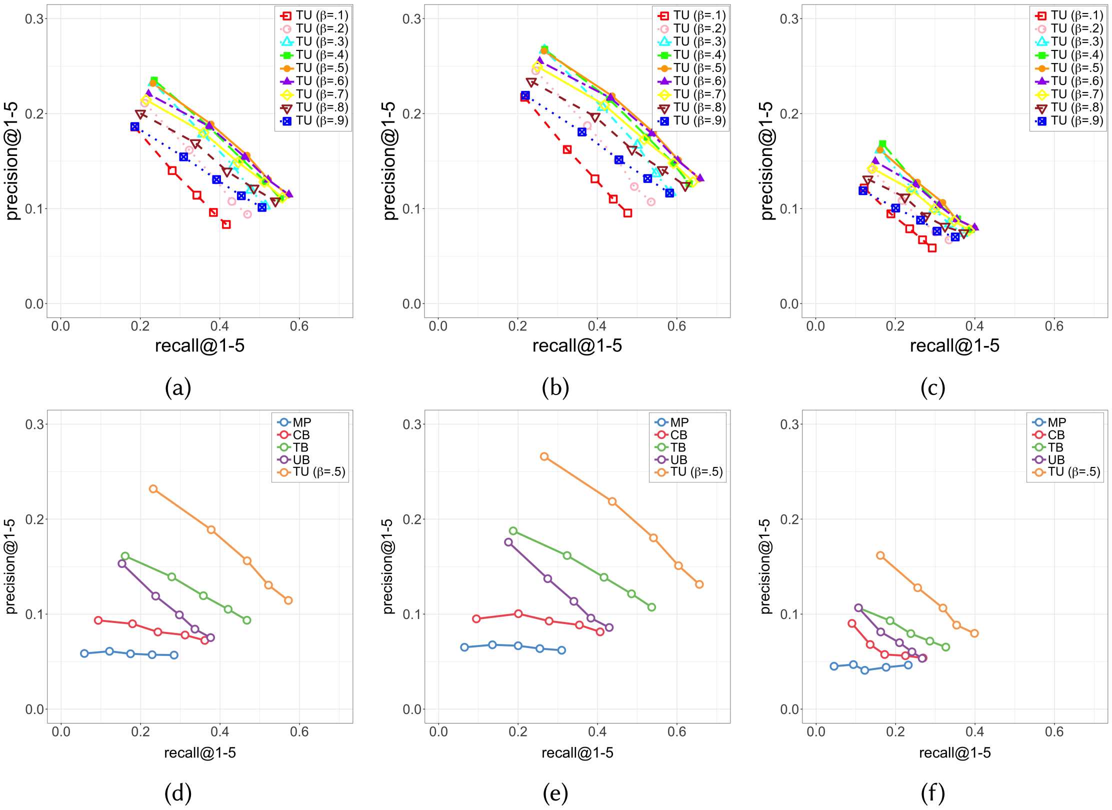
Content-based (CB). It ranks documents according to the relative frequency at which the user-selected terms appear in titles and abstracts. After building a vector space model, it computes document-query similarity as the weighted cosine similarity (see Section 3.3.1).

Most Popular (MP). A topical most-popular model was implemented for this experiment with no personalization features. MP always returns the most frequently chosen documents for the underlying topic (as mentioned in Section 4.2.1).

4.3 Results

This section reports on offline performance assessment and provides further contextual insights. Firstly, the optimal configurations for the TU_β model is identified and then the three variations of social models are contrasted against the content-based (CB) and a most-popular (MP) baselines. Finally, the influence of the underlying search task is investigated. Fig. 5 illustrates all cases with precision vs. recall plots measured $@k = 1, \cdots, 5$.

4.3.1 Optimal Configuration for Tag+User Model. Precision-recall lines in Fig. 5a show that accuracy plummets as β tends to either 0 or 1. A rather steep growth in the ratio can be observed as β grows from .1 to .3, then it stabilizes around $\beta = .5$ and decreases again towards $\beta > .7$. Marginal differences between .4 and .6 are noticeable, but the general trend is that performance reaches its peak towards a balanced weighing of *TB* and *UB* scores. This tendency can also be observed in focused and broad search (Figures 5b and 5c, respectively). In the subsequent analysis, the balanced version of the combined tag+user-based model is taken into account, namely $TU_{\beta=.5}$.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Precision vs. Recall Plots. *First Row*: TU_{β} with $\beta = \{0, \ldots, 1\}$ (step=.1). *Second Row*: Comparisons across *MP*, *CB*, *UB*, *TB* and $TU_{\beta=.5}$. Columns are split task-wise by overall, focused and broad.

4.3.2 Overall Performance. Multiple univariate ANOVA tests for the five metrics listed in Table 2 revealed a significant effect of the model on all performance metrics @5. *F*-statistics matched for classification accuracy (*recall*, *precision* and *F1*) and for ranking accuracy metrics (*nDCG* and *MRR*), given the correlation between metrics of the same kind. *F*-ratios and significance are listed below:

- *Classification Accuracy*: $F(4, 45) = 294$, $p < .001$
- *Ranking Accuracy* $F(4, 45) = 366$, $p < .001$

Pairwise post-hoc Tukey contrasts revealed full ordering at $p < .001$, such that: $TU_{\beta=.5} > TB > UB > CB > MP$. In general, accuracy appears rather low. The characteristics of the ground truth dataset probably played a role in this regard, since validation based on single-item tests can only yield one hit per case. Therefore, we would expect true accuracy to be higher than the measures reported herein. Precision is particularly affected conform k grows because, given a hit, all remaining $k - 1$ items are assumed non relevant. The remainder of this section analyzes in detail the baseline models, followed by comparisons between content-based and social methods. Finally, we discuss which type of social model turned out most effective.

Baseline Algorithms. CB presents better scores than *MP* across all metrics. Precision-recall plot in Fig. 5d shows rather flat lines for both, with *CB* always above *MP*, although the gap between them tends to get closer in each incremental step of k. Pairwise comparisons in Table 2 actually reveal

Table 2. Classification and ranking accuracy across MP, CB, UB, TB and $TU_{\beta=.5}$. @k.

k	**Metric**	MP	CB	UB	TB	$TU_{\beta=.5}$
1	Recall	.059	.093	.153	.161	**.232**
	Precision	.059	.093	.153	.161	**.232**
	F1	.029	.047	.077	.081	**.116**
	nDCG	.059	.093	.153	.161	**.232**
	MRR	.059	.093	.153	.161	**.232**
2	Recall	.122	.180	.238	.278	**.378**
	Precision	.061	.090	.119	.139	**.189**
	F1	.041	.060	.079	.093	**.126**
	nDCG	.098	.148	.207	.235	**.324**
	MRR	.090	.137	.196	.220	**.305**
3	Recall	.175	.244	.298	.358	**.468**
	Precision	.058	.081	.099	.119	**.156**
	F1	.044	.061	.074	.090	**.117**
	nDCG	.125	.180	.237	.275	**.369**
	MRR	.108	.158	.215	.246	**.335**
4	Recall	.229	.312	.337	.420	**.522**
	Precision	.057	.078	.084	.105	**.131**
	F1	.046	.062	.067	.084	**.104**
	nDCG	.148	.209	.253	.302	**.392**
	MRR	.121	.175	.225	.262	**.348**
5	Recall	.284	.362	.376	.468	**.572**
	Precision	.057	.072	.075	.094	**.114**
	F1	.047	.060	.063	.078	**.095**
	nDCG	.170	.228	.269	.320	**.412**
	MRR	.132	.185	.233	.271	**.358**

Note: Values are averaged for focused and broad tasks.

that even though the difference tends to shrink considerably, CB is 60% more accurate @$k = 1$ and 27% @$k = 5$. It is likely that better scores for CB are due to participants bookmarking documents ranked high by the tool in the previous user study.

Content or Social? Fig. 5d shows that UB and TB precision-recall lines start quite close @$k = 1$, with scores roughly 65% higher than CB. Then CB and TB lines remain fairly parallel, while the gap between CB and UB closes abruptly as k is incremented, to the point that the accuracy of UB is only 4% higher than CB's. The line trajectories suggest that with larger top-k results CB would eventually surpass UB.

Regarding ranking accuracy, pairwise comparisons in Table 2 show better outcomes for UB with respect to CB, with a minimum difference of 18% measured for *nDCG*. However, TB still outperforms UB in ranking accuracy, with *nDCG* and *MRR* scores fluctuating between 12 and 19% higher for TB.

Which Social Model? Fig. 5d illustrates an ample difference in recall and precision favoring $TU_{\beta=.5}$. The pure user-based setting UB performs poorly compared to the pure tag-based TB, while the difference between TB and the hybrid $TU_{\beta=.5}$ is even larger. A close look at Table 2 reveals that $TU_{\beta=.5}$ outperforms TB in classification accuracy across all top-k tests by at least 30%. Despite a slightly smaller gap for classification accuracy with k towards 5, the difference in performance remains above 20%. Summarizing the broad picture, the increment in performance from UB to TB across all metrics is roughly 17%, while the overall increment of $TU_{\beta=.5}$ with respect to TB is 33%.

4.3.3 The Influence of Focused and Broad Search. Recall that in the user study from which the bookmark dataset was obtained the subjects performed (per task) (i) two focused searches, whereby 2 or 3 keywords of interest were provided; and (ii) a broad search, where they had to identify search terms upon reading a piece of text. Participants tended to use more keywords with less overlaps for the broad task, hence it was expected that accuracy would fall for broad searches.

Precision-recall plot for broad search task (third row in Fig. 5d) shows an effect that can be visually described as a "compressing accordion", in contrast to the plot for focused search (second row in Fig. 5d). The path for *CB* presents the most abrupt change in its shape. While precision floats around .09 in focused tasks, the effect of the broad search is that it rapidly falls to .06 $@k = 2$ and stays around .05 thereafter. Nevertheless, as it was previously mentioned, *CB* has a foreseeable tendency to match or even outperform UB as k grows, which is already the case under broad task type $@k = 5$. Indeed, it can be observed in the second row of Fig. 5d that CB and UB coincide around the point $(.27, .05)$. *MP* shows a similar tendency. Possibly, further tests $@k > 5$ would reveal that these three models converge. Overall, $TU_{\beta=.5}$ and TB remain as the best and second-best performing approaches, with the gap between their paths still considerably large in broad search.

The loss for *MP* appears less steep in absolute terms, but the relative loss in precision-recall ratio turned out to be of about 30%. Overall precision-recall loss for *CB* was 29%, the lowest across all approaches. UB, TB and $TU_{\beta=.5}$ decrease performance in 39, 42 and 40%, respectively.

Summarizing, the evidence points out that the type of search task affected performance of all the analyzed models, with a consistent decay in broad search with respect to focused search tasks. The social models suffered more than the two baselines, as they became about 10% less accurate than *MP* and *CB* in the case of broad search. Nevertheless. $TU_{\beta=.5}$ remained as the most accurate approach in spite of the greater decline.

5 ONLINE EVALUATION: PERCEIVED ACCURACY

The experiment described in the previous section showed that the social models significantly outperform the content-based and most-popular alternatives, based on a ground-truth dataset of bookmarks collected in the context of exploratory search tasks. The evaluation presented in this section intends to support and extend the findings from the offline assessment, but in an online setup that allows for measuring, at least to some extent, true user satisfaction. In other words, we seek to discover whether the differences found in objective accuracy are also perceived by real users. Moreover, by executing the study in a crowd-sourced platform, it was possible to collect relevance feedback for every item in the ranked list and not just for one known case. Together, the assessment of objective and subjective accuracy are meant to shed light on the individual value of the social models, before integrating them in the social exploratory search system.

5.1 Methodology

With the purpose of collecting first-hand feedback for the models analyzed in the offline evaluation, *MP* and *CB* were again the baseline against which the pure tag-based (*TB*), pure user-based (*UB*) and the hybrid $TU_{\beta=.5}$ were tested. In addition, similarly to the objective assessment, the current evaluation incorporated both focused and broad exploratory search scenarios, whereby participants had to rate documents according to either a couple keywords or a longer text of reference.

Investigating the effect of explanations in users' assessment was of particular interest as well. Given that our approach to social exploratory search aimed to preserve the principle of *transparency* from the original *uRank* system [22], it was important to explore at this early stage how to convey fine-grained relevance information for the different models. Also, explanations could have varying implications for different relevance models or for the two search types. Therefore, at this point we experimented with simple textual explanations with some numeric indicators. Summarizing, the evaluation setup was structured under the following conditions:

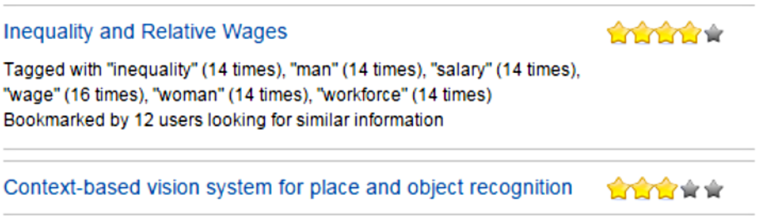
- **Relevance Model**: *MP*, *CB*, *TB*, *UB* and $TU_{\beta=.5}$.
- **Search type**: this study replicated one focused and one broad task from Study I in [22] (Q2 and Q3, respectively).
- **Explanations**: either no explanations or textual. Informative strings for each kind of model were generated, e.g. “Bookmarked 21 times for the topic Augmented Reality” (*MP*), “Frequent terms that appear in this document: gap (moderate), gender (moderate), wage (moderate)” (*CB*), or “Tagged with china (21 times), industrial (21 times), symbiosis (21 times). Bookmarked by 18 users looking for similar information” ($TU_{\beta=.5}$).

5.1.1 Task and Procedure. Participants were only exposed to one model throughout an evaluation session. They were presented with a list of 5 documents at a time, whereby they had to rate all items by relevance according to a text of reference. The subjects had to complete a total of 8 tasks, i.e. each one produced 40 document ratings in total. We prepared four topics, such that every participant worked twice with each topic. The topic, type of search and the presence of explanations were randomly assigned across the 8 tasks, but always avoiding overlaps of equal conditions for the same topic. That is to say, if a participant performed a focused search task with explanations for the topic “robots”, then the next task for that topic would be a broad search without explanations.

Once participants started a session, they were presented with the goal, instructions and a link to the evaluation Web site. The site’s UI included a header at the top with the topic and keywords or sentence of interest, a widget with a list of recommendations, illustrated in Fig. 6, and a juxtaposed section that served as document viewer. Users could access document abstracts by clicking on a title in the list. The instructions given to participants were formulated as follows:

> *Please read carefully the topic and phrase of interest in the page header. Then read document titles and snippets to determine how relevant they are (click on a title in the list to access the snippet). Rate the 5 documents in the list by clicking on the star icons (1 star = not relevant at all, 5 stars = very relevant).*

The same procedure was repeated for the 8 tasks. For each rated item, the system recorded document id, ranking, user rating, and the corresponding conditions, i.e. model, search type and explanations (true or false).



Inequality and Relative Wages ★★★★☆

Tagged with "inequality" (14 times), "man" (14 times), "salary" (14 times), "wage" (16 times), "woman" (14 times), "workforce" (14 times)
Bookmarked by 12 users looking for similar information

Context-based vision system for place and object recognition ★★★☆☆

Fig. 6. Screenshot of rating widget used for online evaluation. (top) Example of a document retrieved with $TU_{\beta=.5}$ for the topic *Women in workforce* presented with textual explanations. (bottom) A document for the topic *Augmented Reality* presented without explanations.

5.1.2 Data Preparation. The selected topics where the same as in the experiment described in Section 4, which allowed us to re-use bookmark data as well. Also it provided a fair ground to draw comparisons between the finding in offline and online settings.

In the case of *MP*, search results were obtained by choosing the most popular documents per topic. In order to compute document-query similarities in *CB*, a vector space model was generated for all documents. In turn, the social models were trained following the same procedure as in the offline experimental setup. It is important to mention that participants were treated as new to the system, thus no user models with past searches were available for the social models. Search result lists were pre-compiled for all algorithms, assuring that participants would have to rate the same items across all methods and topic.

The process to automatically generate the queries varied depending on the type of search task. For focused search, the models received the 2 or 3 keywords defined for Q2 as input, while for broad search tasks, all nouns and adjectives were extracted from the text provided in Q3 and used as model inputs. The text of reference and generated queries are listed in Appendix A.

5.1.3 Participants. The study was published on Prolific[3], a platform facilitating the recruitment of participants for crowd-sourced studies. Invitations were automatically sent out via Email to eligible users registered in the platform. A total of 234 Prolific users took part in the evaluation ($MP = 49,\ CB = 49,\ TB = 50,\ TU_{\beta=.5} = 50,\ UB = 36$[4]), who were paid GBP 1.50 (~USD 2) for an estimated time of 15 minutes. Demographic details are listed below.

- *Gender*: 141 male, 92 female, 1 unspecified.
- *Age*: 63 [< 25], 110 [25 – 34], 41 [35 – 44], 20 [> 44].
- *Nationality*: 101 UK, 35 USA, 16 Serbia, 16 Bosnia and Herzegovina, 14 India, 52 other.
- *Student status*: 74 yes, 157 no, 3 unspecified.

5.1.4 Accuracy Metric. System accuracy was computed in terms of *utility* [12]. This metric calculates a score for the whole list (rather than individual items) based on user ratings. Also, it considers that the worth of a retrieved item declines as it falls in lower positions. Formula 14 indicates how utility is computed for a 5-item list rated by user u.

$$Ut_u = \sum_{j=1}^{5} \frac{max(r_{ui_j} - d,\ 0)}{2^{\frac{j-1}{\alpha-1}}} \tag{14}$$

where r_{ui_j} is the rating given by u to the item in the j^{th} position, i_j, d is a “don’t care” threshold (set to 2 like in [11]) and α is a half-life parameter that corresponds to the position of the item in the list with 50% probability of being inspected. This value was set to 3 in order to obtain maximum utility (all items are actually rated and hence inspected).

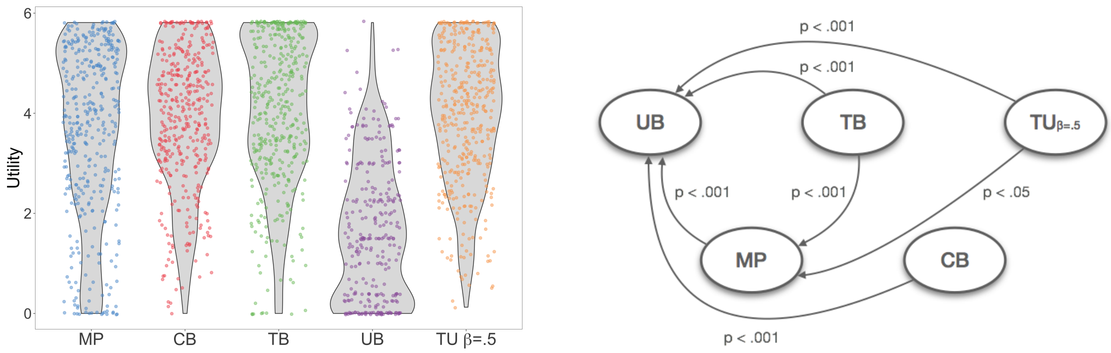
[3] https://prolific.ac

[4] Fewer users for *UB* was not intentional. The study condition remained open on *Prolific* until no more participants took part in it after a week.

Table 3. Utility values broken down by gender and topics.

Gender	WW	Ro	AR	CE	All topics
Female	3.63	3.72	3.77	3.28	3.60
Male	3.37	3.64	3.64	3.19	3.46
Overall	3.47	3.68	3.70	3.23	3.52

Note: WW = Women in Workforce; Ro = Robots; AR = Augmented Reality; CE = Circular Economy



(a) Violin plots showing distribution of utility (0 = no utility, 6= perfect utility).

(b) Partial ordering of multiple pairwise comparisons across the five models.

Fig. 7. Results of overall perceived accuracy in crowd-sourced evaluation of social models.

5.2 Results

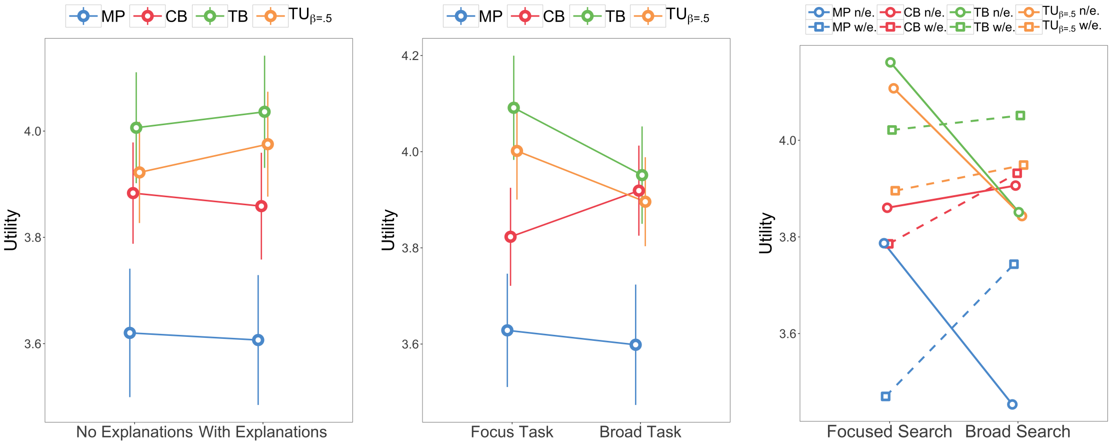
This section reports on system accuracy as perceived by users rating the 5-item lists of documents for the given information needs. Before analyzing performance at relevance model level, we checked for gender biases across topics with a two-way ANOVA test. As it can be observed in Table 3, average utility was marginally higher for female participants, though not significantly, $F(1, 231) = 0.64$, $p = .42$. Conversely, utility across topics was significantly different, $F(1, 1625) = 19.96$, $p < .001$. Nevertheless, the effect of the interaction between gender and topic was not significant, $F(3, 1625) = 0.49$, $p = 69$, which means that scores varied consistently for female and male participants across topics.

In the remainder, we compare overall performance for the five analyzed models and subsequently describe the effects of providing textual explanations and of the different search tasks. Since each user rated 8 lists under all possible combinations of explanation, search type and topic, overall ratings cannot be considered independent because they are clustered per user. Therefore, we conducted the analysis by fitting linear mixed-effects models with user as random effect. Modeling the user as random effect allows for taking into account the within-group variance of clustered responses. We executed ANOVA tests on the mixed-effect models and report F-ratio and p-value. Post-hoc Tukey multiple comparisons across search methods are reported accordingly in Section 5.2.1. Goodness-of-fit is indicated as well in terms of R^2 coefficients[5].

5.2.1 Overall Perceived Accuracy. We first fitted a linear mixed-effects model to estimate the independent effect of search method on utility scores, hence in this case search method was the only independent. The model accounts for 55% of the variance, $R^2 = .55$. R^2 is not an absolute measure of goodness-of-fit. Instead, it should be interpreted in accordance to the context. Given that modeling human behavior is complex, linear models in psychology or social sciences rarely account for more than 50% of the variance. Therefore we can cautiously consider that our model explains a reasonable portion of it.

The model revealed a significant independent effect of search method on total utility, $F(4, 229) = 41.1$, $p < .001$. The violin plots in Fig. 7a illustrate the distribution of utility scores, which allows for discovering at a glance that the overall difference is caused mostly by the poor performance of the user-based model (*UB*). Post-hoc pairwise comparisons showed that none of the search models

[5] Linear mixed-effect models belong to the family of generalized linear mixed models (GLMM), for which the computation of R^2 is not trivial. We used the Ω_0^2 function from Xu [62] for R^2 estimations.



(a) No/with explanations (b) Focused/broad search (c) Explanations and Search

Fig. 8. Interaction lines show the effect of contextual factors: (a) use of textual explanations and (b) type of search task. Standard error bars are shown in (a) and (b), but omitted in (c) to avoid visual clutter. Mean scores for *UB* are below 3 in all cases and henceforth excluded to ease comparisons between the other methods.

outperformed all the others. However, the partial ordering, illustrated in Fig. 7b, indicates that social models that incorporate tag information, namely TB and $TU_{\beta=.5}$ achieved maximal performance among the five tested approaches. The baseline MP ad CB share the intermediate position, while UB was clearly perceived as the worst model.

Except for UB, all other algorithms scored an average utility above 3.6, with TB reaching 4.02. Similarly, user rating showed that on average users were satisfied with the listed items, as the median rating was between 3.8 and 4 for the two tag-based social models and the baselines as well. Again, UB lagged behind with a median rating of only 2.4, clearly below the half-life parameter of utility (α in Formula 14). A possible reason for UB's under-performance is the quasi collaborative filtering nature of the model, specifically in the neighbor-document similarity computation in Formula 6. The target user's neighborhood is formed by other users whose past interests (manifested through bookmarking actions) match the current search. However, given that an average user normally bookmarks documents about more than just one topic over time, UB can also match documents that were not necessarily bookmarked for the current information needs of the target user. In other words, UB tends to produce much more diverse results than TB. Diversification in general serves to bring results with a broader insight, in a way to promote serendipitous findings [53]. Although serendipity is an important component of information-seeking activity [26], particularly for ill-defined problems and exploration of new domains [44, 61], UB's results were likely too diverse and hence perceived as off-topic. Notwithstanding, the reader should not overlook that the balanced fusion of tag and collaborative information in the $TU_{\beta=.5}$ model is indeed perceived as (at least marginally) more accurate than content-based result lists.

5.2.2 The Effect of Textual Explanations. Taking the base linear model, the next step was to add explanations as a second exogenous variable. It turned out that the independent effect of textual explanations on utility was not significant, $F(1, 1633) = 0.89$, $p = .35$. Also, the interaction between search method and explanations did not have a significant effect, $F(4, 1633) = 0.80$, $p = .52$, as it can be observed in the rather parallel lines in Fig. 8a. The extended mixed-effects model again explains 55% of the variance, $R^2 = .55$.

In a few words, the results indicate that all search models performed uniformly regardless of the presence of textual explanations. The social models TB and $T_{\beta=.5}$ did however show a marginal growth when explanations were added, while the moderately negative slope for CB indicates that it suffered the opposite effect.

5.2.3 Focused and Broad Search. Similarly to the case of explanations, we added search task as a second variable to the base mixed-effects model (with search method as first independent variable). Neither the independent effect of search task nor the interaction with search method were significant. Interestingly, although differences between CB and the two tag-based models were not significant, Fig. 8b reveals that TB and $T_{\beta=.5}$ perform above CB in focused tasks and then tend to converge in broad search tasks.

In a next step, we partitioned the data and used only records for focused task to fit the base model, i.e. taking search model as the only independent variable. The effect of search model was still significant in the case of focused search, $F(4, 229) = 34$, $p < .001$. Then as expected, UB performed poorly against the other four models ($p < .001$ in all post-hoc comparisons). It was also found that the gap between the social models and the most-popular baseline is significant, with $p < .05$ for TB and $p < .05$ (one-sided) for $TU_{\beta=.5}$.

Following the same procedure for the broad search data points, we observed that search method maintained a significant effect on perceived accuracy, $F(4, 229) = 39$, $p < .001$, although differences between tag-based social models and most-popular disappear.

5.2.4 Combined Effect of Search Task and Explanations. The two previous sections (5.2.2 and 5.2.3) reported that the independent effect of textual explanations and search task were not significant. Overall, the interaction between the two is not significant either (regardless of search method).

Nevertheless, the search methods seem to behave differently when the transitions between no/with explanation and focused/broad search tasks are analyzed all together. Individual reactions can be observed in Fig. 8c, where solid and dashed lines of the same colors cross each other. In all cases of focused tasks, documents presented without explanations were rated higher than with explanations. This is denoted as the origin of full lines (no explanations) is always higher than for dashed lines (with explanations) of the same color. Conversely, this tendency is reversed in broad searches, where utility scores appear higher when textual explanations are present rather than in their absence.

In order to analyze this "crossing" effect at individual levels, we partitioned the data in five subsets, one per search method. Subsequently, we executed independent tests by fitting five multilevel models[6], each with one subset, and computed the effect of the interaction between explanations and search task. The interaction effect was not significant for any group, except for MP, $F(1, 340) = 5$, $p = .03$. MP's performance decreases substantially with explanations in focused tasks and without explanations in broad searches (lower blue points on the left and right side of Fig. 8c, respectively). This suggests that when users have very specific search needs, explanations claiming that a document is amongst the most popular for certain topic are counterproductive. Conversely, topical explanations seem to be helpful when users deal with less defined information needs.

All in all, these results denote that CB and the social models keep a rather stable performance irrespective of search task and the presence of textual explanations. Notwithstanding, despite only marginal differences, explanations appear more useful in the case of broad searches, where information needs are ill-defined in contrast to focused searches.

[6] Type I error corrections are not applied because the multilevel models do not test differences across relevance methods, but instead at intra-group level, i.e. separately for the five search models.

6 DISCUSSION ON OBJECTIVE AND PERCEIVED ACCURACY

The evaluations presented in the last two sections reported on objective and perceived performance for the social relevance models introduced in section 3.3. These two evaluations provided insight that was useful at design stages to determine whether adding social relevance models to an exploratory search system could bring potential benefits. Summarizing the evaluation process, we set a most-popular and a content-based model as baselines and then measured their performance according to several accuracy metrics. For a comprehensive analysis, we also analyzed the influence of factors such as the type of search task and the use of textual explanations. As a result, the collected evidence allowed us to set an informed judgment of the potential value of pursuing user-controllable social exploratory search. Noteworthy, keeping the performance assessment at this stage separate from the actual system also allowed us to avoid biases caused by how users perceive system characteristics through the user interface. The remainder of this section discusses the contrasting findings from the offline and online evaluations and our decision to include the two social models in the resulting hybrid *uRank* system (Section 3).

The evaluation described in section 4 was executed offline with data collected in a previous user study. The objective assessment of performance suggests that:

- Social approaches are able to incorporate information needs to tackle the item prediction problem for exploratory search more effectively than popularity and content-based methods.
- A hybrid scoring scheme between *TB* and *UB* is stronger than the two models alone, whereby the balanced fusion $TU_{\beta=.5}$ appears as the most adequate blend.
- The type of search task had a general impact on all models, with the social ones showing larger decays from focused to broad search. Pure tag-based and combined tag+user models still outperformed the social user-based and the baselines.

Thereafter, the crowd-sourced subjective evaluation reported in section 5 revealed interesting findings, some of them contradicting the objective assessment:

- Pure user-based social search (UB) turned out unsuitable for the information-seeking task. However, blended with tagging information significantly improved its perceived accuracy.
- Social models that incorporate tagging information showed maximal performance, either in its pure version (TB) and balanced fusion ($TU_{\beta=.5}$).
- The use of textual explanations and the type of search alone did not influence perceived accuracy. However, there is a tendency towards higher accuracy without explanations in focused search tasks, whereas in broad search explanations seem more effective.

There is a conspicuous mismatch between objective and subjective measurements of system accuracy, which beyond deeming a pure user-based approach unfit, indicates that perceived differences between tag-based social models and the content-based baseline are not as large as they seemed in the offline experiment. Nevertheless this is not discouraging but the opposite. In the context of learning to rank search results, it is not the idea to entirely replace usual content-based approaches. After all, users pursue an information-seeking task where the content of retrieved items cannot be overlooked for the sake of purely collaborative or tagging approaches. Applying social models alone can work well for a different kind of information access paradigm, for example when users receive recommendations of documents via email or through an application like *Conference Navigator* [13], where a recommeder matches user profiles to conference talks. Conversely, given the nature of exploratory search, leveraging traces left by other users to provide an implicit kind of collaboration seems to be more suitable for a scenario in which the user “forages” search results as usual and, in addition, social cues enable them to identify resources that were likely useful for similar information needs in the past.

All in all, the evidence collected via objective and subjective performance assessments indicates that incorporating social models into exploratory search systems appears as a promising path to improve system accuracy and the user experience. Based on the outcomes observed in social search models, particularly the good performance of the balanced model with user and tag information $TU_{\beta=.5}$, we sustain our design decision of incorporating both, the TB and UB models into the hybrid search system.

Nevertheless, it should be noted that the two previous evaluation have their limitations. Firstly, the baseline models are not the state-of-the-art in information retrieval or recommender systems. Nonetheless, we considered them appropriate, in particular CB, given that it proved useful in the context of user-driven exploratory search tasks [21, 22]. Supporting this kind of task is, ultimately, the purpose of incorporating social models to a controllable UI (Section 3.1). Secondly, the nature of the dataset could be somewhat biased. We attempted to produce a balanced dataset by working with four randomized topics in the original study [21]. However, the logged data comes from user behavior, which could not be considered completely unbiased. We cannot dismiss the possibility that the observed results would vary with a different dataset. For example, the optimal configuration for the TU_β model could give more prominence to one of the two social models (β closer to 0 or 1) rather than settle for the balanced fusion thereof. Moreover, the offline assessment is not able to capture true user feedback, but rather test predictive models. As for the online assessment, human raters did not express their own evolving information needs. These were instead simulated with broad and focused searches. Naturally, these limitations respond to the need of collecting a substantial amount of data for a quantitative analysis, thus our constraint to a fixed set of information needs. Therefore, the final study of our evaluation roadmap, presented in the next section, evaluates the full-fledged hybrid *uRank* system, whereby users conducted a more realistic exploratory search task and were in control of the fusion between content and social sources of relevance.

7 USER STUDY: USER-CONTROLLABLE SOCIAL EXPLORATORY SEARCH

We conducted an online user study to assess the worth of *user-controllable social exploratory search* (SES) in contrast to pure ES. Note that we do not aim to compare SES against basic systems without support for exploration. Instead we assess social exploratory search when users are aware and able to control it, compared to an ES system that already outperforms a traditional list-based UI (c.f., [21]). While developing the hybrid tool with numerous controllable and explanatory features, we expected these additions to positively influence system accuracy and user experience. However, the complexity of the UI could result in a system that is too difficult to use and understand. In this study we address such concerns in detail.

7.1 Evaluation Methodology

The study used a between-subject design with conditions ES and SES. As the social search system requires training data to produce document scores, we split the study execution in two stages:

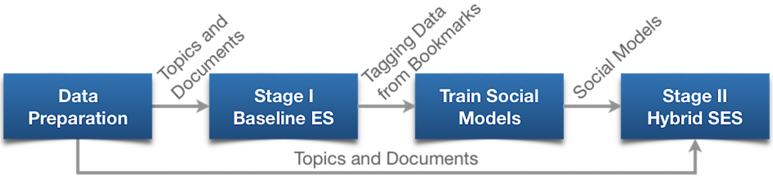
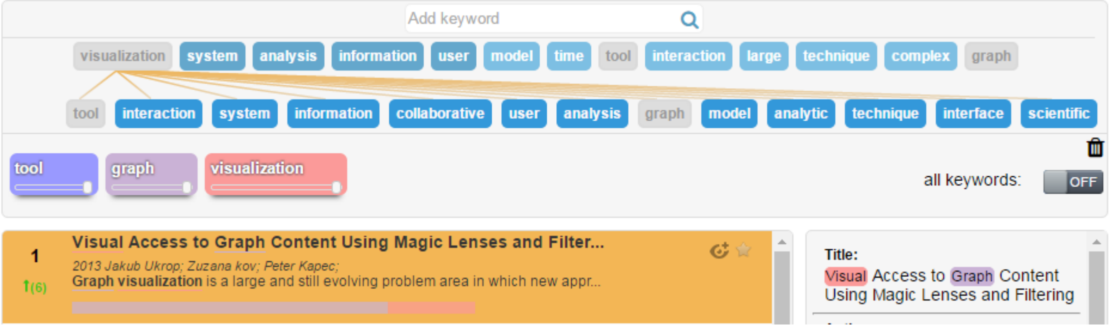


Fig. 9. Stages in User Study

(i) with baseline ES condition, (ii) with hybrid SES condition, after training social models with bookmarks collected in (i). Participants of both conditions performed the same task. Figure 9 illustrates the sequence of executed steps.

7.1.1 Baseline System. For the ES condition, we opted for a system that already supports exploratory search, but lacks social search capabilities. We used the base version of our system with the ranking computed solely on the CB model. The interface in ES resembles that of the hybrid system, as shown in Figure 10, except for the absence of sliders to control the hybrid model and explanatory hints and tooltip described in Section 3.1.2. As a result the baseline system includes most features of the hybrid system, but is inherently less complex. In other words, the user only needs to care about selecting keywords to refine information needs, without concerning about model parameters.



Visual Access to Graph Content Using Magic Lenses and Filter...

2013 Jakub Ukrop; Zuzana kov; Peter Kapec;

Graph visualization is a large and still evolving problem area in which new appr...

Title:
Visual Access to Graph Content Using Magic Lenses and Filtering

Fig. 10. UI of baseline system employed in the user study. Documents are ranked solely by the CB method.

7.1.2 Data Preparation. The need to collect training data imposed a constraint on the number of topics used in the study. Therefore we chose 5 topics from the Computer Science field: *Augmented Reality*, *Visual Analytics*, *Recommender Systems*, *Deep Learning* and *Human-Computer Interaction*. With the help of three researchers with experience in the chosen topics, we prepared topic descriptions by selecting 2 or 3 paragraphs from known literature and removed all references from the text. We then generated datasets with a script that performed several queries to Mendeley's API[7] and manually added other documents suggested by the experts. Each dataset contained over 800 document surrogates.

After the first stage concluded, we had collected 1395 bookmarks. We further enlarged the training pool by asking the experts to perform the same tasks as regular participants, though without filling the survey. Before starting the second stage, we trained the tag- and user-based models with the generated tagging data.

7.1.3 Task and Procedure. First, participants completed a step-by-step tutorial introducing the main features of the system. They had to perform interactions, e.g. keyword selection and weighting, adjusting ranking model parameters, bookmarking and collection management. The tutorial also covered explanatory features such as color-coded bars and ranking-type icons (in hybrid UI).

After the tutorial, participants were presented with a view prompting them to sort the five topics by familiarity. The least familiar topic (i.e. at the bottom) was automatically picked. The task consisted in organizing and collecting resources for the given topic. We asked participants to imagine they had to write an essay about the assigned topic. We provided the corresponding

[7] http://dev.mendeley.com/

topic description, which served as the introduction section. Participants had to read it and find relevant concepts that helped them define subsequent sections of the fictitious essay. Then they had to work with the assigned tool and find relevant documents for each section. To fulfill the task, participants had to create at least 3 collections (one per section) and bookmark at least 5 documents per collection, i.e. at least 15 in total. Upon completion, the participant had to click on "Done" and fill a survey for subjective feedback.

7.1.4 Participants. We recruited participants through the crowd-sourced platform Prolific[8] and from mailing lists of colleagues in the field. In the case of Prolific, the eligibility requirement was that candidates had to hold at least a Bachelor degree in Computer Science. Since the tasks required dealing with scientific content, it was crucial to address users with the appropriate background. As for the field of expertise, the intention was to match the topics covered in the prepared datasets. A total of 79 people took part in the study, from which 43 worked with the baseline system and the remaining 36 with the hybrid UI. 10 participants were recruited via e-mail, while the other 69 completed the study on Prolific. Demographic information is summarized as follows:

- *Age*: 1 $[< 20]$ years old, 66 $[20-29]$, 12 $[30-39]$.
- *Gender*: 62 male, 17 female.
- *Country of birth*: India (18), USA (15), UK (9), other (37).
- *Highest level of education*: 39 Bachelor (BSc), 23 Master (MSc or similar), 17 other.
- *Proficiency in English language*: 6 basic, 9 intermediate, 30 advanced, 34 native speaker.
- *Familiarity with search interfaces, e.g. Google* (-3 =not at all familiar, 3 =very familiar): $mean = 2.65$, $sd = 0.78$.

7.1.5 Measurements.

User Behavior. The two systems collected action logs for keyword manipulations, document clicks, bookmarks (and their position in the list). For the hybrid tool, we also logged manipulations of model parameter sliders.

System Accuracy. For each user u, we computed *average Discounted Cumulative Gain* (*aDCG*) as the mean *DCG* value across all u's bookmarks, B_u, similarly to the *nDCG* calculation in Section 4.2.2.

$$aDCG = \frac{1}{|B_u|} \sum_{d \in B_u} \frac{2^{rel_d} - 1}{\log(rank_d + 1)} \tag{15}$$

The reason to employ this metric instead of the popular normalized version is that *nDCG* was conceived for a typical retrieval problem, where it is possible to estimate an ideal ranking and measure the accumulated gain from the top to the bottom of a result list for a single query. In our scenario, it can be argued that subsequent interactions with keyword tags affect the original query and, hence, we are dealing with a session rather than single queries. Although another metric called session DCG (*sDCG*) [35] is able to compute performance in a multi-query session, it was thought for the usual query-response paradigm, where formulating new queries is costly. Thus the more query refinement is need, the more penalization *sDCG* adds. We consider that adding a penalization for selecting many keyword tags is not suitable, precisely because our interactive tool seeks to promote this behavior. Also, it is not clear whether a simple selection should be taken as a whole new query. In further consequence, we opted to simply measure *DCG* and compute the average to compensate for the fact that users could bookmark an uneven number of documents.

[8] www.prolific.ac

UX Scale for Subjective Feedback. Instead of using a standard but too generic questionnaire, like SUS [42], we created a custom scale to measure the particular aspects of the user experience (UX) targeted in this evaluation. The custom questionnaire included several items addressing specific latent constructs. Latent constructs are often difficult to measure in one question because they can have disparate interpretations. For example, the concept of "perceived control" belongs in the jargon of research on user experience but can be difficult to interpret for lay users. Therefore, this kind of constructs are better inferred from a number of observed indicators, e.g. survey questions. The survey comprised 28 questions covering perceived system aspects, user satisfaction and personal characteristics. Answers were recorded on a 7-point Likert scale (-3 = strongly disagree, 3 = strongly agree), with some questions worded in negative tone (reversed for the analysis).

7.1.6 Modeling the User Experience. This study focused on causal and mediated effects in system accuracy, user behavior and subjective feedback all together. Therefore, we conducted the statistical analysis using structural equation modeling techniques. We built the model shown in Figure 11 in a two-step approach [5]: first we fitted a measurement model (to calculate factor loadings) and then a structural model (measurement model plus causal relationships, see Section 7.1.7). UX concepts and directionality of causal effects are partly grounded on the framework proposed by Knijnenburg et al. [40] to model the user experience with recommender systems.

At first, we planned for 5 latent constructs: 3 subjective system aspects (*SSA*) (perceived control, transparency and result quality) and 2 experience (*EXP*) (choice satisfaction and satisfaction with the system). "Choice satisfaction" and "satisfaction with the system" had poor discriminant validity (inter-factor correlation = .97), i.e. users did not distinguish from one another and perceived the two constructs as the same one. Since they both measure "satisfaction" aspects, we merged them into a single factor[9]. We excluded a total of 8 survey items from the analysis due to low communality or high cross-loading. In particular, all 4 items for "perceived result quality" had to be excluded as 2 of them had low communality[10]. The final three factors listed below showed good convergent validity, internal consistency reliability and discriminant validity. Appendix B provides full disclosure of items and validation of the UX scale.

- *Perceived Control* (SSA) 3 items, e.g. "The system allowed me to easily refine my search terms".
- *Perceived Transparency* (SSA) 3 items, e.g. "The system gave me a sense of transparency".
- *Satisfaction* (EXP) 9 items, e.g. "I'm satisfied with the documents I bookmarked".
- The item "Are you familiar with search user interfaces?" is a single personal characteristic (PC) observed indicator.

Confirmatory factor analysis with a mean- and variance-adjusted weighted least squares estimator yielded admissible model fit[11], $\chi^2(87) = 105.68$, $p = .08$; $CFI = .99$; $TLI = .99$; $RMSEA = .05$ with 90% $CI = [0, .08]$.

7.1.7 Hypotheses in Structural Model. Intuitively, the hybrid system requires several fine-tuning actions to adjust for information needs accordingly. We then expected to observe differences in system performance and user satisfaction when users actively controlled the system via keyword interactions ($\#kw.int$ for short) and weighing of model parameters. Also, the only difference in control features between the two systems is the absence of model weight sliders in the baseline. Clearly, only those participants that manipulated the ranking weights in the hybrid tool were

[9] In practice, some respecification of the measurement model is necessary. It is acceptable as long as it is not grounded in statistical considerations alone but in conjunction with theory and content [5].

[10] Using less than 3 indicators per factor is discouraged.

[11] χ^2 test should be non-significant. Acceptable cut-off values: $CFI > .96$, $TLI > .95$, $RMSEA < .05$ with 90% CI upper-bound $< .10$

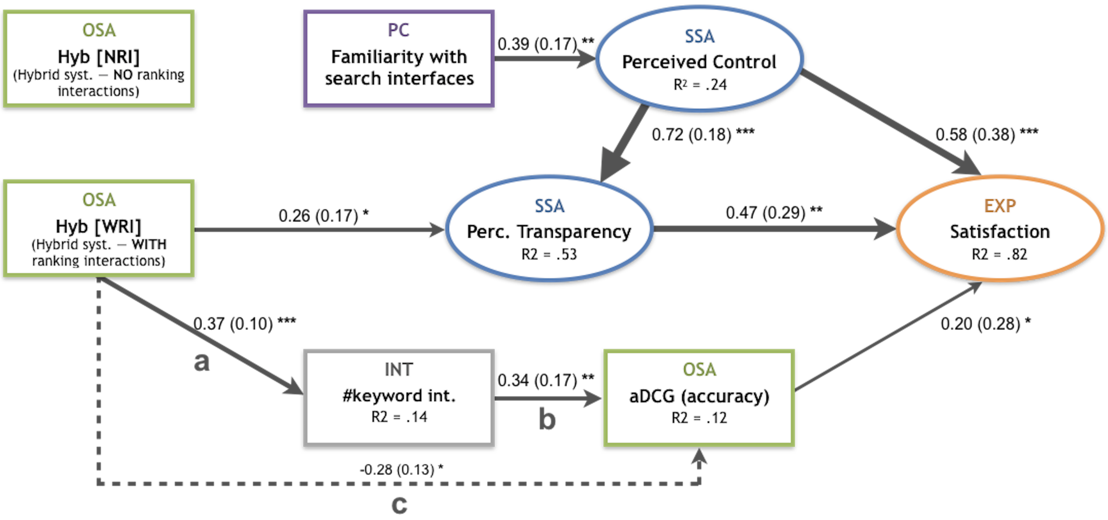


Fig. 11. Structural model of system accuracy and user experience. Model Elements: Objective (*OSA*) and Subjective System Aspects (*SSA*), Personal Characteristics (*PC*), Interactions (*INT*) and Experience (*EXP*). Pathways represent significant causal relationships, with numbers and thickness indicating regression coefficients (with robust SE). Factors are scaled with variances set to 1. Full standardized solution is reported , i.e. path coefficients indicate increments in SD units. Full (dashed) lines indicate significant positive (negative) coefficients. Significance: $^{***}\ p < .001$, $^{**}\ p < .01$, $^{*}\ p < .05$.

able to “appreciate” any difference with respect to the baseline (not consciously though due to the between-subject setup).

Consequently, we split participants of the hybrid system into 2 sub-groups: participants that performed at least one interaction with the ranking model, or “ranking interactions” for short, (Hyb_{WRI}, $N = 22$) and those that did not (Hyb_{NRI}, $N = 14$). Participants in the first group accumulated 5.05 ranking interactions on average ($SD = 5.19$, 17 of them with more than one). The two sub-groups represent manipulated variations, noted as objective system aspects (*OSA*) in the model diagram. Pathways originating from these nodes indicate comparisons *against the baseline* ($N = 43$). We then built the structural model in Figure 11 based on the following hypotheses:

H1: The hybrid system will be significantly more accurate than the baseline (*OSA*). As we highlighted the importance of interactions for the system to understand user needs and thus be more accurate, we hypothesized higher system performance for the hybrid tool with ranking interactions, though influenced (at least to some extent) by the amount of keyword interactions ($Hyb_{WRI} \rightarrow \#kw.int \rightarrow aDCG$, where $\#kw.int$ is an observed behavior or interaction (*INT*)).

H2: Participants will perceive higher degree of control and transparency with the hybrid system (*SSA*). The hybrid system provides features for enhanced control (model sliders in Figure 1.C) and transparency (tooltip and hint icons in Figure 4). At least users of the Hyb_{WRI} sub-group should perceive more control over the system and transparency provided by graphic explanations. Although at first we assumed that perceived result quality would mediate the effect of perceived control and transparency on satisfaction ($Perc.Control + Perc.Transp \rightarrow Perc.Res.Quality \rightarrow Satisfaction$), the fact that the latter had to be excluded from the model does not prevent us from measuring the direct effect ($Perc.Control + Perc.Transp \rightarrow Satisfaction$).

H3: Expertise will influence users' perception of the system *(PC).* Personal characteristics of a user are known to influence their experience with a system, but cannot be accounted for by the system itself [16, 38, 39]. As such, our assumption is that, regardless of the tool employed by a user, their level of familiarity with search interfaces will contribute to how controllable and transparent they perceive the system.

H4: The hybrid system will produce higher user satisfaction (*EXP*). Ultimately, we expected that objective (*OSA*) and subjective aspects (*SSA*) of the system would lead to higher user satisfaction (*EXP*). Although the direct link between system accuracy (*OSA*) and user experience (*EXP*) is often not evident or weak, Knijnenburg et al.'s framework [40] allows to model user experience as a result of how the user perceives the characteristics of the system ($OSA \rightarrow SSA \rightarrow EXP$).

7.2 Results

We applied structural equation modeling to analyze causal and mediated effects in behavioral and subjective data all together. The model in Figure 11 is the cornerstone to validate our hypotheses with respect to performance and user experience. The model obtained excellent goodness-of-fit, $\chi^2(141) = 146.55$, $p = .36$, $CFI = .998$, $TLI = .998$, $RMSEA = .022$ with 90% $CI = [0, .058]$. Regression pathways report standardized coefficients (β) and robust standard errors.

7.2.1 System Performance. We found no difference in performance (*aDCG*) between the subgroup of hybrid system users with no ranking interactions and the baseline condition, $\beta = 0.06$, $SE = 0.11$, ns (denoted by the absence of links originating from the Hyb_{NRI} node). Conversely, when hybrid system participants interacted with ranking sliders, they performed significantly more keyword interactions ($\#kw.int$), as it can be observed in the density distribution in Figure 12. In turn, more keyword interactions had a positive effect on system accuracy. Regression paths denote two types of effect of Hyb_{WRI} on $aDCG$: *(i)* a direct negative effect (path c in Figure 11), and *(ii)* an indirect positive effect through $\#kw.int$ (pathways a and b). Coefficients for indirect and direct effects with opposite signs indicate the possibility that the mediator ($\#kw.int$) acts as a suppressor.

To measure the amount of mediation of Hyb_{WRI} on $aDCG$ due to $\#kw.int$, we first computed estimates for the indirect effect (product of a and b) and the total effect ($c + a * b$). Then we applied bootstrap[12] with 20,000 draws to obtain standard errors and 95% confidence intervals. We found a significant indirect effect, $\beta = 0.11$, $SE = 0.060$, $CI = [0.030, 0.277]$ (CI does not include 0, i.e. $p < .05$), while the total effect turned out non significant, $\beta = -0.16$, $SE = 0.106$, $CI = [-0.368, 0.051]$. A significant indirect effect that reduces the total effect to 0 indicates complete mediation. In other words, pathway c connecting Hyb_{WRI} with $aDCG$ in Figure 11 is broken, and thus the entire effect of the independent variable (Hyb_{WRI}) on the dependent variable ($aDCG$) is transmitted through the mediator ($\#kw.int$). This evidence supports **H1**.

Considering the caveat of careless interactions, we corroborated that $\#kw.int$ correlates with overall session time ($r = .5$, $p < .001$). Participants of the hybrid system with full interactions opened more document abstracts per bookmark (0.85 vs. 0.54), although not statistically significant ($U = 400$, $p = .08$). For visual assessment, Figure 13 plots the temporal progression of $aDCG$ scores (with SE intervals) for the first 15 minutes of an average session of the 3 groups (baseline, Hyb_{NRI} and Hyb_{WRI}). Mean accuracy does not differ significantly, but it remains higher for Hyb_{WRI} throughout the average session time ($M = 10'39''$, $SD = 5'53''$). To sum up, the evidence suggests that the hybrid system outperforms the baseline, *provided that users actively make use of control features*, i.e. interactions with keyword tags and ranking model sliders.

[12] Bootstrap is a non-parametric technique based on resampling with replacement, often used for testing significance of indirect and total effects in mediation analysis [58].

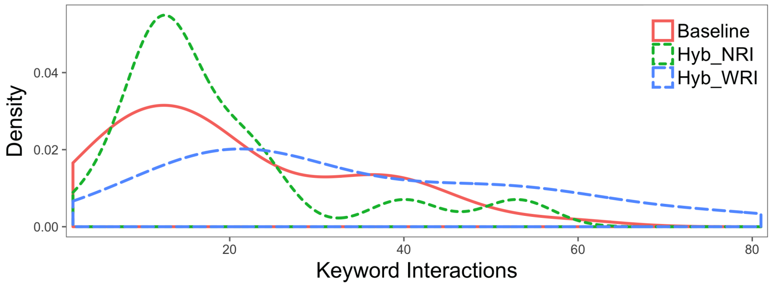


Fig. 12. Density distribution of keyword interactions for the 3 groups analyzed in the user study.

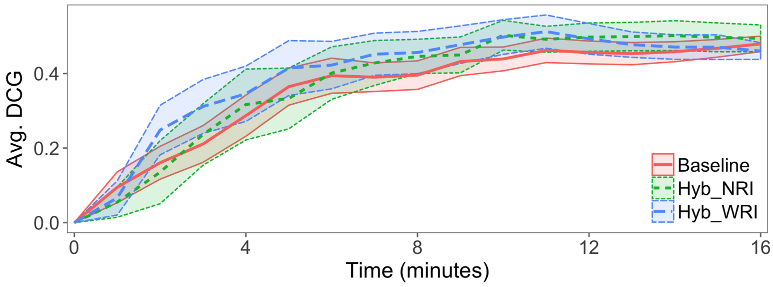


Fig. 13. Temporal evolution of cumulative *aDCG*.

7.2.2 Perceived Control and Transparency. Participants did not perceive a significant difference in the *level of control* between the two systems, which could be attributed to the baseline having similar features to interactively control the document ranking (except for ranking sliders). In turn, the hybrid tool with ranking interactions (Hyb_{WRI}) had a direct positive effect on users' perception of *transparency*. Although both systems convey keyword contribution to overall scores (through stacked-colored bars), only the hybrid one provides explanations about the methods used to compute the ranking. These results partially support **H2**. In short, the SES system influenced the perception of transparency but not of control. Moreover, we found no difference between the baseline and participants without ranking interactions (Hyb_{NRI}), which supports our initial assumption that the user should fully leverage control features of the hybrid system to appreciate the benefit of combining social and exploratory search.

7.2.3 Familiarity with Search Interfaces. Regardless of the used system, participants acknowledging high familiarity with search interfaces perceived a greater level of control. As it has been observed in previous research, users who understand a system also perceive more control over it [38]. Thus, it is likely that more experienced users understood these features and learned how to apply them and manipulate the system with less effort. The model also reveals a large effect of perceived control on transparency. More specifically, perceiving a system as transparent seems to be greatly influenced by how controllable it is regarded by the user and indirectly by their search experience. This provides evidence of the impact of personal characteristics on the traits that users are able to recognize in a system, as stated in **H3**.

Note that other studies have already investigated the effects of search experience and expertise (which are not necessarily the same [8]) in information-seeking contexts, reporting better task outcomes [7, 36], lower task completion time [55] and more effective strategies [46, 49]. All these studies focused on some aspect of performance. Conversely, we analyzed the effect of self-assessed experience on perceived system traits and user satisfaction, as a proxy to potential adoption of an advanced search system. Notwithstanding, it is worth mentioning that this personal characteristic did not have an effect on system accuracy in the model, hence the lack of absence of a pathway from the *PC* node ("familiarity with search interfaces") to *aDCG* (accuracy) in Figure 11.

7.2.4 User Satisfaction. Although sometimes the effect of system accuracy cannot be directly assessed in the experience (*EXP*), but through subjective perception of system aspects [16], in this case we found evidence of a direct effect. We modeled "satisfaction" as the final output of multiple objective and subjective variables. From the statistical evidence we can corroborate that higher satisfaction is a result of not only a good-performing system but also interaction and presentation aspects in the UI. Perceived control and transparency had each a positive effect on user satisfaction. Familiarity with search interfaces contributes to user satisfaction because experienced participants felt more control over the system. This can be linked to previous findings reporting that domain experts are more satisfied when they can control the system [39]. It is important to highlight that the effect of the hybrid tool for the Hyb_{WRI} group on user satisfaction is not direct but transferable through perceived transparency (*SSA*) and system accuracy (*OSA*) (mediated by keyword interactions), thus supporting **H4**.

7.3 Discussion and Limitations

In this study we assessed system performance and user experience for a social exploratory search system contrasting it with a system supporting exploratory search alone. Allowing the user to control not only keywords (a rather familiar action for any Web user), but also parameters of a hybrid ranking model, entails a higher level of complexity. Nevertheless, an important contribution of this study is that social exploratory search achieved better system performance, as long as users exploited its full potential to refine both, keywords of interest and ranking parameters. Another key finding is that system accuracy contributed directly to higher user satisfaction.

As for perceived system traits, people who felt more in control also understood better the system's logic ($Perc. Control \rightarrow Perc. Transparency$). That is to say, the system is able to explain itself by way of interactions and visual explanations. However, the model attributes a higher perception of control to users' personal traits rather than to the incremental customization in the hybrid system. The positive effect of perceived control and transparency is directly transferred to user satisfaction. This result aligns to research in social recommender systems [11, 38], though it had not been previously analyzed in exploratory search contexts.

It is worth noting that, despite higher complexity, the hybrid system was still as engaging as the exploratory search baseline. Users in SES condition that did not adjust the hybrid model (fusion of content and social features), perceived similar levels of control and transparency, and were as satisfied as the baseline users. This is denoted by the absence of pathways originating from Hyb_{NRI} in the structural model. In other words, the extra complexity was not detrimental to user experience. We expect that novice users would be able to learn to exploit the advanced features with better initial guidance and after spending more time (beyond a single session).

As for shortcomings, the online nature of our study is at the same time an advantage and a limitation. On one hand, we believe the physical presence of an observer can lead to (intuitively positive) opinion bias. Moreover, through the crowd-sourced platform we managed to recruit more participants than we would have been able in a controlled in-lab study. On the other hand, we lacked the opportunity to make our own observations and interview the participants. The participants' background in Computer Science can also be considered a limitation, arguably unavoidable since the social and collaborative models needed as much training data as possible on related topics. Generalizing to all backgrounds is not trivial but from our past experience with students from life sciences [21] and the influence of experience with search interfaces, we expect that any user habituated to research tasks should be able to understand the system after some trial and error.

Finally, we need to acknowledge that the proposed model is one among infinite possibilities. This limitation is extrinsic and in general extends to the structural equation modeling methodology. Our hypotheses, although grounded in existing theories and common sense, reflect our own conceptualization of the user experience in a specific context. Thus, we do not claim that our model is absolute but a valid alternative.

8 CONCLUSIONS AND FUTURE WORK

In this paper we presented the series of evaluations conducted to validate our approach to user-controllable social exploratory search, published in [20]. An exploratory search system was extended with social search capabilities by leveraging information left by past users, in this case bookmarking behavior. Social features were implemented in term of two social relevance models, one based on social tags and collaborative user matching. In turn, control and transparency, important features in recommender systems but less exploited in exploratory search systems, allowed for fusing the social component in a consistent way. On one hand, the user is empowered to shape their information needs and decide to what extent social information should be taken into account and, on the other hand, the system effectively communicates the effect of the user's decisions.

Two evaluations conducted prior to effective extension of the hybrid *uRank* system allowed us to determine the potential value of two models of social relevance. The offline and crowd-sourced experiments contrasted objective and perceived accuracy, in experimental setups that were independent from the user interface and, hence, enabled us to observe their performance without biases due to presentation or interaction aspects. The offline experiment revealed that the two social models (and the balanced fusion of them) significantly outperforms a content-based and a most-popular models. However, these outcomes were strongly disputed by online crowd-sourced ratings. The user matching method yielded very poor perceived accuracy, in comparison to the tag-based and content-based models. Nevertheless, the balanced merge between tag and user models ($TU_{\beta=.5}$) had similar subjective performance as the pure tag-based model. This final outcome suggested that the two models are strong working together. Thereby, we incorporated them both into the extended hybrid interface.

Subsequently, an empirical user study reported evidence of the benefits of user-controllable fusion of content and social components for exploratory search tasks, in terms of system performance and user experience. Based on the hypotheses delineated in a structural regressions model, we corroborated that social exploratory search produces more accurate results, as long as the user actively "tells" the system how their information needs evolve and the right blend of content and social information. In turn, the system was perceived as transparent due to the explanatory nature of the UI. Conversely, the perception of control is mostly attributed to the users' own traits rather than differences among the two systems. Together, system accuracy and perceived control and transparency promoted higher user satisfaction.

As a final remark on methodology, we should mention that our experimental roadmap did not follow a strictly delineated plan, but it is rather the result of cumulative experience throughout the three phases (offline, online evaluations and user study) and previous research. It is known that repeatability of HCI studies is not trivial. Several researchers have addressed this issue in recent workshops and seminars [6, 25]. In our effort to facilitate reproducibility, this paper described in as much detail as possible the procedure for data collection, task preparation, questionnaires, data analysis, etc. We believe the incremental evaluations here presented add to a robust research outcome. It is difficult to assess, however, whether the overall process could be simplified or shortened. In our humble contribution, we list the steps that could serve other researchers to conduct user-centered studies (particularly for exploratory search, but also applicable to e.g. RS):

(1) Collect pertinent data from previous studies (own or third-party research). Check for relevant tasks.
(2) Test offline. Choose baselines and compute accuracy with a few state-of-the-art metrics.
(3) Repeat evaluation setup (i.e. tested conditions), but in an o this time with users (in neutral UI). Keep simple task and UI, as well as data collection and metrics, e.g. compute utility based on ratings.
(4) Incorporate successful models into UI and back-end.
(5) Test full-fledged system against suitable baseline. Points to bear in mind are: prepare a realistic task, collect behavior logs, create or reuse suitable survey for subjective feedback, and apply advanced statistical analysis to exploit gathered data, e.g. CFA and SEM techniques.

Regarding the source on information feeding the social approach in this work, it consists entirely of bookmark data, which is of an implicit nature, i.e users do not explicitly rate or provide any kind of relevance feedback. Other sources could be explored as well, especially since not all users might leverage the bookmaking mechanism. An alternative implicit source could be, for example, a logsonomy [41]. Logsonomies build folksonomy structures based on click-log data. In our case, we could assume that when a user clicks (inspects) a document, the user considers that document to be relevant for the currently selected keywords. An advantage of this approach is that it would allow us to produce larger volumes of relevance judgments. A clear disadvantage is that opening a document does not necessarily mean that the user deems it relevant, but rather a candidate or a promising resource. Hence, the quality of relevance judgments based on logsonomy would be intuitively inferior to bookmark data. A path of interest in our future work is to extend bookmark-based relationships with click-log information and subsequently learn weights for each source type.

Another important challenge for the near future is to further exploit the interactive features to learn task models and improve ranking accuracy. Also, an open question is whether simpler or fewer features would produce similar results in system performance and user experience. More importantly, our next goal is to investigate whether an adaptive system that personalizes the amount of advanced features exposed to a particular user (based on expertise and background) has similar effects to the ones observed in this work. Finally, a longitudinal evaluation is necessary to validate the observed results in a multi-session context, for example with students performing search-to-learn activities in technology-enhanced learning platforms or intelligent tutors.

APPENDIX

A QUERIES IN ONLINE EVALUATION

This section lists text of interest provided to participants of the online evaluation described in Section 5 (in italics), along with the query terms submitted to the analyzed models (in parentheses). Items (1) and (2) correspond to focused tasks, while (3) indicates broad tasks.

Women in Workforce (WW):

(1) *Participation of women in the workforce* (woman, workforce, participation)
(2) *Gender wage gap* (gender, wage, gap)
(3) *Women in the workforce earning wages or a salary are part of a modern phenomenon, one that developed at the same time as the growth of paid employment for men; yet women have been challenged by inequality in the workforce.* (e.g. woman, workforce, wage, salary, man, inequality)

Robots (Ro):

(1) *Autonomous robots* (autonomous, robots)

(2) *Human-robot interaction* (human, robot, interaction)

(3) *The branch of technology that deals with the design, construction, operation, and application of robots, as well as computer systems for their control, sensory feedback, and information processing is robotics.* (robot, control, sensor, information, processing)

Augmented Reality (AR):

(1) *Virtual environments* (virtual, environment)

(2) *Context-based objection recognition* (context, object, recognition)

(3) *Augmented reality (AR) is a live direct or indirect view of a physical, real-world environment whose elements are augmented (or supplemented) by computer-generated sensory input such as sound, video, graphics or GPS data.* (augmented, environment, image, reality, video, world)

Circular Economy (CE):

(1) *Waste management* (waste, management)

(2) *Industrial symbiosis in China* (industrial, symbiosis, china)

(3) *A circular economy naturally encompasses a shift from fossil fuels to the use of renewable energy, the eradication of waste and the role of diversity as a characteristic of resilient and productive systems.* (circular, economy, fossil, fuel, system, waste)

Table 4. User experience scale for user study in Section 7

Factor	#	Question	Loading	Variance
Perceived Result Quality (SSA)	rq1	The system provided good results		
	rq2	The system provided too many bad results (rev)		
	rq3	Items at the top of the list were normally relevant for my chosen keywords		
	rq4	Good documents were hard to find (rev)		
Perceived Control (SSA)	co1	I felt the system had limited functionalities (rev)	0.86	0.26
	co2	The system was useful to find good combinations of keywords	0.88	0.24
	co3	The system allowed me to refine my search terms easily	0.76	0.43
AVE = .69	co4	I felt that I was able to tell the system exactly what I wanted it to do		
α = .82	co5	The components of the system were hard to handle (rev)		
Perceived Transparency (SSA)	tr1	It was easy to understand why some documents were ranked higher than others		
	tr2	I found the system very intuitive	0.83	0.31
	tr3	The system gave me a sense of transparency	0.70	0.52
AVE = .66	tr4	I feel the user interface was self-explanatory		
α = .82	tr5	The system showed a lot of consistency	0.90	0.19
Satisfaction (EXP)	cs1	I'm satisfied with the documents I bookmarked	0.71	0.50
	cs2	If I had to write an essay about the given topics, I think I would use my bookmarks as references	0.78	0.39
AVE = .69 α = .93	cs3	If a colleague/friend asked me for information about the topic, I would probably suggest my bookmarked documents	0.89	0.21
	cs4	I know other documents that are better than the ones I bookmarked (rev)	0.87	0.24
	ss1	The system was helpful to understand the given topic	0.75	0.43
	ss2	I see a potential benefit in using the system for research tasks	0.80	0.37
	ss3	I felt confident using the system	0.85	0.29
	ss4	I would recommend the system to a friend/colleague	0.74	0.46
	ss5	Working with the system was exhausting (-)	0.87	0.24
Search Exp.	se1	Are you familiar with search user interfaces?	-	-

Note: AVE indicates convergent validity. Cut-off value is .50. Cronbach's α indicates internal consistency reliability, $\alpha > .8$ is good, $> .9$ is excellent. Inter-factor correlations should be $< .85$ for discriminant validity.

B USER EXPERIENCE SCALE

The user study in Section 7 applied a survey consisting of a 28-item custom scale modeling the user experience. The scale comprised 4 factors: (i) two subjective system aspects (SSA), *perceived control* and *perceived transparency*; (ii) one experience (EXP) factor, *satisfaction*; and (iii) one personal characteristic (PC), *search experience*. The latter was a single indicator and therefore not treated as a factor (factors should comprise at least three questions). Note that factors are abstractions for a group of questions that is not universal, but rather the experimenters' interpretation of a particular construct.

Table 4 shows the factors and loading items. *Perceived result quality* had to be excluded because two if its items had low communality. Items indicated as "cs" and "ss" (second column) in *Satisfaction* correspond to merged items from *Choice Satisfaction* and *Satisfaction with the system*, respectively. Items in gray were excluded due to low communality (high variance) or high cross-loadings (items that significantly loaded on more than one factor). Factor fit is indicated in the left-most column, below factor names. The final three factors showed good convergent validity (average variance extracted – AVE), internal consistency reliability (Cronbach's α) and discriminant validity. The two right-most columns indicate item loadings to factors and their uniqueness, measured as residual variance.

REFERENCES

[1] J.-w. Ahn and P. Brusilovsky. 2013. Adaptive visualization for exploratory information retrieval. *Inf. Process. Manag.* 49, 5 (2013), 1139–1164.

[2] J.-w. Ahn, P. Brusilovsky, J. Grady, D. He, and R. Florian. 2010. Semantic annotation based exploratory search for information analysts. *Inf. Process. Manag.* 46, 4 (2010), 383–402.

[3] J.-w. Ahn, P. Brusilovsky, D. He, J. Grady, and Q. Li. 2008. Personalized web exploration with task models. *Proc. WWW '08* (2008), 1–10.

[4] J.-w. Ahn, R. Farzan, and P. Brusilovsky. 2006. Social search in the context of social navigation. *Journal of the Korean Society for Information Management* 23, 2 (2006), 147–165.

[5] J. C. Anderson and D. W. Gerbing. 1988. Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach. *Psychological Bulletin* 103, 3 (5 1988), 411–423.

[6] S. Arabas, M. R. Bareford, L. R. de Silva, I. P. Gent, B. M. Gorman, M. Hajiarabderkani, T. Henderson, L. Hutton, A. Konovalov, L. Kotthoff, et al. 2014. Case Studies and Challenges in Reproducibility in the Computational Sciences. *arXiv preprint arXiv:1408.2123* (2014).

[7] A. Aula. 2005. *Studying user strategies and characteristics for developing web search interfaces.* Ph.D. Dissertation. University of Tampere.

[8] A. Aula and K. Nordhausen. 2006. Modeling successful performance in Web searching. *Journal of the American Society for Information Science and Technology* 57, 12 (oct 2006), 1678–1693.

[9] R. Baeza-Yates, C. Hurtado, and M. Mendoza. 2004. Query Recommendation Using Query Logs in Search Engines. In *Proc. EDBT '04*. Springer-Verlag, 588–596.

[10] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. 2007. Optimizing web search using social annotations. In *Proc. WWW '07*. ACM, 501–510.

[11] S. Bostandjiev, J. O'Donovan, and T. Höllerer. 2012. TasteWeights: A Visual Interactive Hybrid Recommender System. In *Proc. RecSys '12*. ACM, 35–42.

[12] J. S. Breese, D. Heckerman, and C. Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. *Proc. Uncertainty in Artificial Intelligence* (1998), 43–52.

[13] P. Brusilovsky, J. S. Oh, C. López, D. Parra, and W. Jeng. 2017. Linking information and people in a social system for academic conferences. *New Review of Hypermedia and Multimedia* 23, 2 (2017), 81–111.

[14] P. Brusilovsky, B. Smyth, and B. Shapira. 2018. *Social Search.* Springer International Publishing, 213–276.

[15] R. Burke. 2002. Hybrid Recommender Systems: Survey and Experiments. *User Model. User-adapt. Interact.* 12, 4 (Nov. 2002), 331–370.

[16] D. N. Chin. 2001. Empirical Evaluation of User Models and User-Adapted Systems. *User Model. User-adapt. Interact.* 11, 1-2 (March 2001), 181–194.

[17] J. Chuang, C. D. Manning, and J. Heer. 2012. "Without the Clutter of Unimportant Words": Descriptive Keyphrases for Text Visualization. *ACM Trans. Comput.-Hum. Interact.* 19, 3, Article 19 (Oct. 2012), 29 pages.

[18] N. Craswell. 2009. Encyclopedia of Database Systems. Springer US, Boston, MA, Chapter Mean Recip, 1703.

[19] P. Cremonesi, Y. Koren, and R. Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proc. RecSys '10.* 39—-46.

[20] C. di Sciascio, P. Brusilovsky, and E. Veas. 2018. A Study on User-Controllable Social Exploratory Search. In *Proc. IUI '18.* ACM, 353–364.

[21] C. di Sciascio, V. Sabol, and E. Veas. 2016. Rank As You Go : User-Driven Exploration of Search Results. In *Proc. IUI '16 (IUI '16).* ACM, 118–129.

[22] C. di Sciascio, V. Sabol, and E. Veas. 2017. Supporting Exploratory Search with a Visual User-Driven Approach. *ACM Transactions on Interactive Intelligent Systems* 7, 4 (Dec 2017), 1–35.

[23] M. D. Ekstrand, D. Kluver, F. M. Harper, and J. A. Konstan. 2015. Letting Users Choose Recommender Algorithms : An Experimental Study. In *Proceedings of the 9th ACM Conference on Recommender systems - RecSys '15.* 11–18.

[24] B. M. Evans and E. H. Chi. 2010. An elaborated model of social search. *Inf. Process. Manag.* 46, 6 (2010), 656–678.

[25] N. Ferro, N. Kando, N. Fuhr, M. Lippold, and J. Kalervo. 2016. Increasing Reproducibility in IR : Findings from the Dagstuhl Seminar on "Reproducibility of Data-Oriented Experiments in e-Science" Overview of the PRIMAD Model. *SIGIR Forum* 50, 1 (2016), 68–82.

[26] A. Foster and N. Ford. 2003. Serendipity and information seeking: An empirical study. *Journal of Documentation* 59, 3 (2003), 321–340.

[27] J. Freyne and B. Smyth. 2004. An experiment in social search. In *Adaptive Hypermedia and Adaptive Web-Based Systems.* Springer, 95–103.

[28] E. Gomez-Nieto, F. San Roman, P. Pagliosa, W. Casaca, E. S. Helou, M. C. F. de Oliveira, and L. G. Nonato. 2014. Similarity preserving snippet-based visualization of web search results. *IEEE Trans. Vis. Comput. Graph.* 20, 3 (March 2014), 457–70.

[29] C. He, D. Parra, and K. Verbert. 2016. Interactive Recommender Systems. *Expert Syst. Appl.* 56, C (Sept. 2016), 9–27.

[30] M. A. Hearst. 1995. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proc CHI '95.* ACM Press, 59–66.

[31] M. A. Hearst. 2009. *Search User Interfaces* (1st ed.). Cambridge University Press, New York, NY, USA.

[32] J. L. Herlocker, J. A. Konstan, and J. Riedl. 2000. Explaining collaborative filtering recommendations. In *Proc. CSCW '00.* 241–250.

[33] O. Hoeber and X. D. Yang. 2008. Evaluating WordBars in exploratory Web search scenarios. *Inf. Process. Manag.* 44, 2 (2008), 485–510.

[34] K. Järvelin and J. Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (oct 2002), 422–446.

[35] K. Järvelin, S. L. Price, L. M. L. Delcambre, and M. L. Nielsen. 2008. Discounted cumulated gain based evaluation of multiple-query IR sessions. In *European Conference on Information Retrieval.* Springer, 4–15.

[36] C. Jenkins, C. L. Corritore, and S. Wiedenbeck. 2003. Patterns of information seeking on the Web: A qualitative study of domain expertise and Web expertise. *IT & society* 1, 3 (2003), 64–89.

[37] T. Joachims and F. Radlinski. 2007. Search Engines That Learn from Implicit Feedback. *Computer* 40, 8 (Aug. 2007), 34–40.

[38] B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. 2012. Inspectability and control in social recommenders. In *Proc. RecSys '12.* 43.

[39] B. P. Knijnenburg, N. J. M. Reijmer, and M. C. Willemsen. 2011. Each to His Own: How Different Users Call for Different Interaction Methods in Recommender Systems. In *Proc. RecSys '11.* ACM, 141–148.

[40] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. 2012. Explaining the user experience of recommender systems. *User Modelling and User-Adapted Interaction* 22, 4-5 (2012), 441–504.

[41] B. Krause, R. Jäschke, A. Hotho, and G. Stumme. 2008. Logsonomy - social information retrieval with logdata. In *Proc. HT '08.* 157.

[42] J. R. Lewis and J. Sauro. 2009. The factor structure of the system usability scale. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5619 LNCS (2009), 94–103.

[43] G. Marchionini. 2006. Exploratory Search: From Finding to Understanding. *Commun. ACM* 49, 4 (April 2006), 41–46.

[44] G. Marchionini and B. Shneiderman. 1988. Finding facts vs. browsing knowledge in hypertext systems. *Computer* 21, 1 (1988), 70–80.

[45] A. Micarelli, F. Gasparetti, F. Sciarrone, and S. Gauch. 2007. The Adaptive Web. Springer-Verlag, Berlin, Heidelberg, Chapter Personalized Search on the World Wide Web, 195–230.

[46] R. Navarro-Prieto, M. Scaife, and Y. Rogers. 1999. Cognitive strategies in web searching. In *5th Conference on Human Factors & the Web.* 43–56.

[47] T. Nguyen and J. Zhang. 2006. A Novel Visualization Model for Web Search Results. *IEEE Transactions on Visualization and Computer Graphics* 12, 5 (Sept. 2006), 981–988.

[48] K. A. Olsen, R. R. Korfhage, K. M. Sochats, M. B. Spring, and J. G. Williams. 1993. Visualization of a document collection: The vibe system. *Inf. Process. Manag.* 29, 1 (1993), 69–81.

[49] R. A. Palmquist and K.-S. Kim. 2000. Cognitive style and on-line database search experience as predictors of Web search performance. *Journal of the American Society for Information Science* 51, 6 (2000), 558–566.

[50] D. Parra, P. Brusilovsky, and C. Trattner. 2014. See What You Want to See: Visual User-driven Approach for Hybrid Recommendation. In *Proc. ACM IUI*. ACM, 235–240.

[51] J. Pickens, G. Golovchinsky, C. Shah, P. Qvarfordt, and M. Back. 2008. Algorithmic Mediation for Collaborative Exploratory Search. In *Proc. SIGIR '08*. ACM, 315–322.

[52] P. Pirolli and S. Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. (2005), 2–4.

[53] T. Ruotsalo, G. Jacucci, P. Myllymäki, and S. Kaski. 2015. Interactive Intent Modeling: Information Discovery Beyond Search. *Commun. ACM* 58, 1 (2015), 86–92.

[54] T. Ruotsalo, J. Peltonen, M. Eugster, D. Głowacka, K. Konyushkova, K. Athukorala, I. Kosunen, A. Reijonen, P. Myllymäki, G. Jacucci, and S. Kaski. 2013. Directing exploratory search with interactive intent modeling. In *Proc. ACM CIKM*. 1759—-1764.

[55] H. Saito and K. Miwa. 2001. A cognitive study of information seeking processes in the WWW: the effects of searcher's knowledge and experience. In *Proc. Web Information Systems Engineering*. IEEE Comput. Soc, 321–327.

[56] G. Shani and N. Tractinsky. 2013. Displaying Relevance Scores for Search Results. In *Proc. ACM SIGIR*. ACM, 901–904.

[57] B. Shneiderman, D. Byrd, and W. B. Croft. 1998. Sorting out Searching: A User-interface Framework for Text Searches. *Commun. ACM* 41, 4 (April 1998), 95–98.

[58] P. E. Shrout and N. Bolger. 2002. Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods* 7, 4 (2002), 422–445.

[59] K. Sparck Jones. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 28, 1 (jan 1972), 11–21.

[60] R. W. White, B. Kules, S. M. Drucker, et al. 2006. Supporting exploratory search. *Commun. ACM* 49, 4 (2006), 36–39.

[61] M. L. Wilson, B. Kules, M. C. Schraefel, and B. Shneiderman. 2010. From Keyword Search to Exploration: Designing Future Search Interfaces for the Web. *Found. Trends Web Sci.* 2, 1 (Jan. 2010), 1–97.

[62] R. Xu. 2003. Measuring explained variation in linear mixed effects models. *Statistics in Medicine* 22, 22 (2003), 3527–3541.

[63] K.-P. Yee, K. Swearingen, K. Li, and M. A. Hearst. 2003. Faceted metadata for image search and browsing. In *Proc. CHI '03*. ACM Press, 401–408.

[64] Z. Yue and D. He. 2018. *Collaborative Information Search*. Springer International Publishing, Cham, 108–141.

Received February 2007; revised March 2009; accepted June 2009