Signpost from the Masses: Learning Effects in an Exploratory Social Tag Search Browser

Yvonne Kammerer[+], Rowan Nairn, Peter Pirolli, Ed H. Chi

[+]Knowledge Media Research Center
Konrad-Adenauer-Str. 40
D-72072 Tubingen, Germany
y.kammerer@iwm-kmrc.de

Palo Alto Research Center
Augmented Social Cognition Group
3333 Coyote Hill Road, Palo Alto, CA 94304 USA
{rnairn, pirolli, echi}@parc.com

ABSTRACT

Social tagging arose out of the need to organize found content that is worth revisiting. A significant side effect has been the use of social tagging sites as navigational signposts for interesting content. The collective behavior of users who tagged contents seems to offer a good basis for exploratory search interfaces, even for users who are not using social bookmarking sites. In this paper, we present the design of a tag-based exploratory system and detail an experiment in understanding its effectiveness. The tag-based search system allows users to utilize relevance feedback on tags to indicate their interest in various topics, enabling rapid exploration of the topic space. The experiment shows that the system seems to provide a kind of scaffold for users to learn new topics.

Author Keywords
Social Tagging, Exploratory Interfaces, Social Search

ACM Classification Keywords
H3.3 [Information Search and Retrieval]: Relevance Feedback, Search Process, Selection Process; H5.2. [Information interfaces and presentation]: User Interfaces

INTRODUCTION

Social tagging (or social bookmarking) has increasingly become a common method for users to store, organize, and share labeled bookmarks to online content. Often the tagging is for personal use [11] but a substantial number of people use shared or publically available bookmarks to explore and find information. As noted by Millen et al. [15] social tagging systems provide a mix of direct and indirect "navigational advice" based on the collective behavior of those who have already tagged and organized content. Therefore, social tagging systems seem to be a good basis for *exploratory search* capabilities.

As outlined by Marchionini [14], exploratory search involves ill-structured problems and more open-ended goals, with persistent, opportunistic, iterative, multi-faceted processes aimed more at learning than answering a specific query. Whereas for the fact-retrieval searches, an optimal path to the document(s) containing the required information is crucial, learning and investigation activities lead to a more continuous and exploratory process with the knowledge acquired during this "journey" being essential as well [19]. Therefore, the aim of our tag search browser is to support users' exploratory search by presenting related tags (apart from the results list) and providing the opportunity for relevance feedback.

The design of our exploratory search system is based on social tagging data we obtained by crawling the Web. The problem with freeform social tagging sites is that, as the systems grow, their information signal declines and noise increases, due to synonyms, misspellings, and other linguistic morphologies [3]. We have designed a system that aims to perform a tag normalization that reduces the noise and finds the patterns of co-occurrence between tags to offer a kind of recommendation of related tags and contents. The related tags help deal with the vocabulary problem during search [7]. These recommendations offer support to the user while exploring an unfamiliar topic area.

In this paper, we present the interaction and UI design of the tag search browser called MrTaggy, and an experimental analysis of some learning effects in this exploratory tag search browser. One aim is to evaluate the browser itself to understand its capabilities. Another aim is to demonstrate some learning assessment methods that might prove useful in evaluations of other exploratory search tools such as faceted browsing and searching systems.

RELATED WORK

Vannevar Bush's vision of the Memex [2] has inspired the evolution of information systems that augment and enhance human abilities to find, store, organize, understand, retrieve, and share knowledge. The areas of information retrieval, personal information management, and the Web (to name just a few) have for the most part, historically been focused on supporting *individual* information foraging and sensemaking.

1

Recently there has been an efflorescence of systems aimed at supporting *social* information foraging and sensemaking. These include social tagging and bookmarking systems for photos (e.g., flickr.com), videos (e.g., youtube.com), or Web pages (e.g., del.icio.us). Tagging systems provide a means for users to generate labeled links to content that, at a later time, can be browsed and searched. A unique aspect of tagging systems is the freedom that users have in choosing the vocabulary used to tag objects: any free-form keyword is allowed as a tag. Tags can be organized to provide meaningful navigation structures, and, consequently, can be viewed as an external representation of what the users learned from a page and of how they chose to organize that knowledge.

Several researchers in CSCW have noted how bookmarks and tags serve as signals to other in the community. Lee found that analyses of del.icio.us users who perceive greater degrees of social presence are more likely to annotate their bookmarks to facilitate sharing and discovery [13]. Golder and Huberman's study also showed that there is remarkable regularity in the structure of the social tagging systems that is suggestive of a productive peer-to-peer knowledge system [9].

Researchers in the HCI community have noted the similarity of the cognitive processes between keyword generation during tagging by individual users and the keyword generation during search [6]. The generation of keywords during search is also known as the "vocabulary problem" [7]. Many researchers in the information retrieval community have already explored the use of query logs for aiding later searchers [16, 4, 8]. For example, Glance showed how past queries can be effectively mined to suggest related queries to others [8].

Using social tagging data as "navigational advice" and suggestions for additional vocabulary terms, we are interested in designing exploratory search systems that could help novice users gain knowledge in a topic area more quickly. However, social tagging data generate a vast amount of noise in the forms of synonyms, other linguistic morphologies, and deliberate spam [3]. Previous research shows that an information theoretic analysis of tag usage in del.icio.us bookmarks is suggestive of decreased efficiency in using tags as navigational aids [3].

We have designed a system that enables users to quickly give relevance feedbacks to the system to narrow down to related concepts and relevant URLs. The idea here is to bootstrap the user quickly with other related concepts that might be gleamed from social usage of related tags. Moreover, the popularities of various URLs are suggestive of the best information sources to consult.

In this paper, we will first briefly describe the design and user interaction model of the system, and then detail an experimental study of the overall system, particularly focusing on whether the system helps bootstrap users in unfamiliar topic domains and a learning effect assessment of the exploratory search mechanisms.

MRTAGGY: TAG-BASED SEARCH BROWSER

The tag search browser MrTaggy uses social tagging data to recommend and search through documents by using the relationships between tags and documents to suggest other tags and documents.

Figure 1 shows a typical view of the tag search browser. MrTaggy provides explicit search capabilities (search box and search results list) combined with relevance feedback [1, 17] for query refinements. Users have the opportunity to give relevance feedback to the system in two different ways:

Related Page Feedback: By clicking on the downward arrow a search result can be excluded from the results list whereas by clicking on the upward arrow the search result can be emphasized which leads to an emphasis of other similar Web pages.

Related Tag Feedback: At the left of the user interface a *related tags list* is presented (see Figure 1), which is an
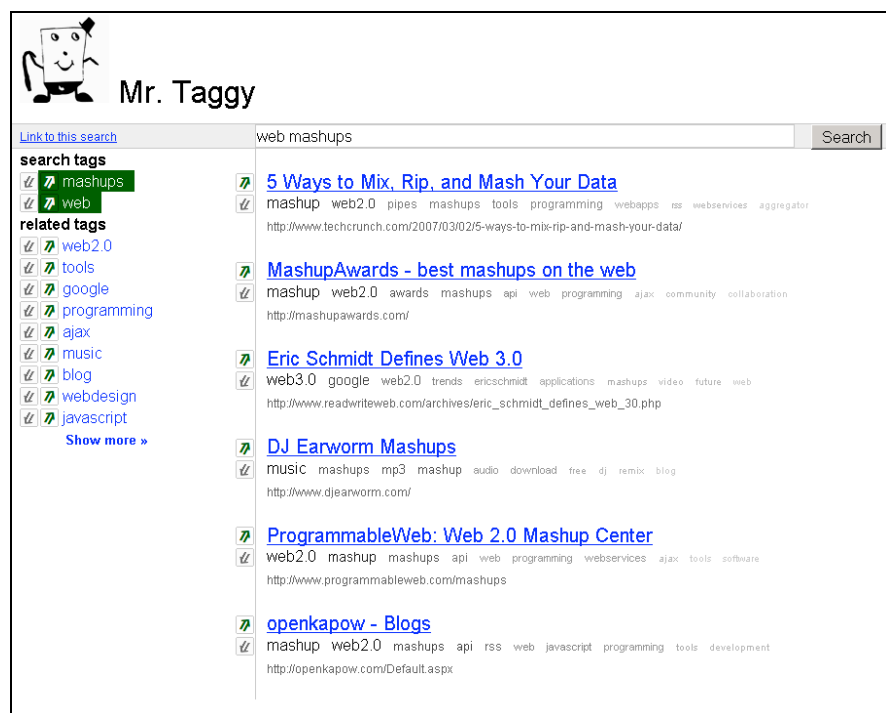

Figure 1. MrTaggy user interface with related tags list on the left and search results lists presented on the right.

overview of other tags related to the keywords typed into the search box. For each related tag, up and down arrows are displayed to enable the user to give relevance feedbacks. The arrows here can be used for query refinements either by adding a relevant tag or by excluding an irrelevant one (see Figure 2).

In addition, users can refine the search result using tags associated with each of the search results. During search, result snippets (see Figure 3) are displayed in the search results list. In addition to the title and the URL of the corresponding Web page, instead of a short summary description, a series of tags are displayed. These tags are applied by other users to label the corresponding Web page. When hovering over tags presented in the snippet, up and down arrows are displayed to enable relevance feedbacks on these tags as well.

Users' relevance feedback actions lead to an immediate reordering or filtering of the results list, since the relevance feedback and the search result list are tightly coupled in the interface. We use animations to display the reordering of the search results, which emphasizes the

Figure 2. MrTaggy user interface with "search tags" section for added tags and "bad tags" section for excluded tags (both on the left).

Figure 3. The 3 parts of a search result snippet in the MrTaggy interface: title, tags, URL.

changes that occurred in the result list (see Video). New search results due to the refinements are marked with a yellow stripe.

A BRIEF DESCRIPTION OF TAGSEARCH ALGORITHM

Having just described the interaction of the relevance feedback part of the system, we now describe how it operates in concert with the backend. Figure 4 shows an architecture diagram of the overall system.

First, a crawling module goes out to the Web and crawls social tagging sites, looking for tuples of the form <User, URL, Tag, Time>. The tuples are kept track of in a MySQL database. In our current system, we have roughly 120 million tuples.

A MapReduce system based on Bayesian inference and spreading activation then computes the probability of each URL or tag being relevant given a particular combination of other tags and URLs. Here we first construct a bigraph between URL and tags based on the tuples and then precompute spreading activation patterns across the graph.
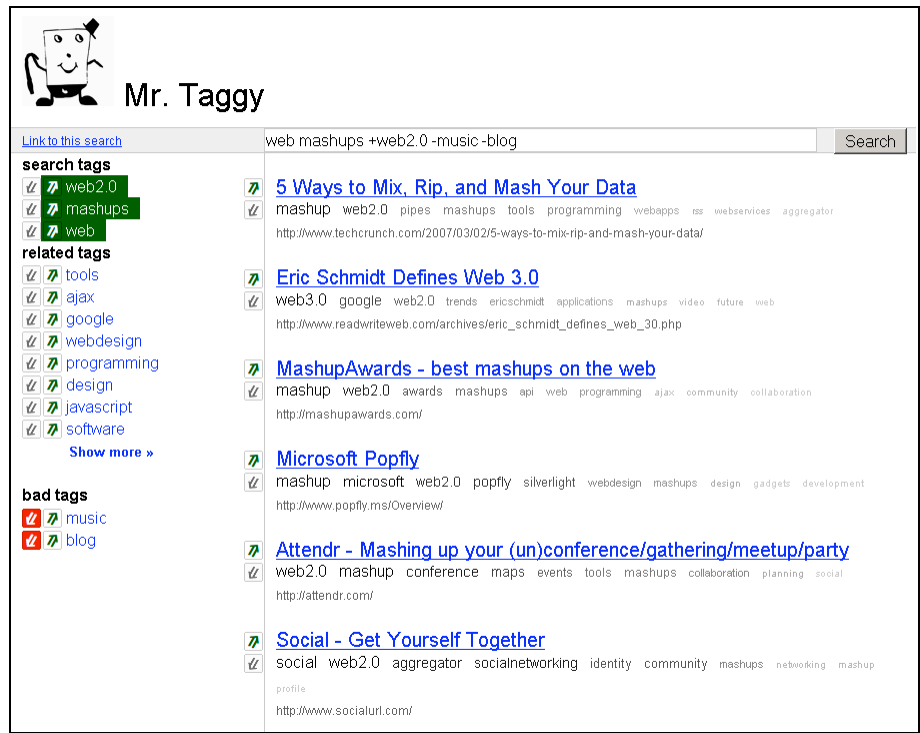
To do this backend computation in massively parallel way, we used the MapReduce framework provided by Hadoop (hadoop.apache org). The results are stored in a Lucene index (lucene.apache.org) so that we can make the retrieval of spreading activation patterns as fast as possible.

Finally, a Web server serves up the search results along with an interactive frontend. The frontend responds to user interaction with relevance feedback arrows by communicating with the Web server using AJAX techniques and animating the interface to an updated state.

In terms of data flow, when the user first issues a query, the Web server looks up the related tag recommendations as well as the URL recommendations in the Lucene index and returns the results back to the frontend client. The client presents the result to the users with the arrows buttons as relevance feedback mechanisms. When the user presses on one of the arrow buttons, the client issues an updated query to the Web server, and a new result set is returned to the client.
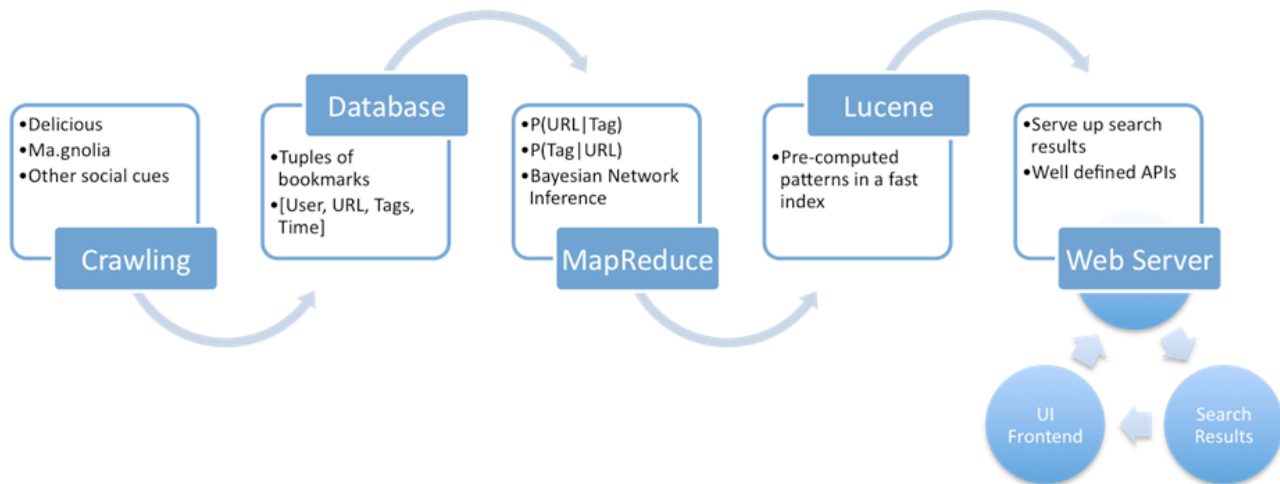
3

Figure 4. Overall architectural diagram of the MrTaggy tag-based search browser.

AN EXPERIMENTAL ANALYSIS OF MRTAGGY AS AN EXPLORATORY SEARCH SYSTEM

As noted above, exploratory search is construed as the product of ill-structured information-seeking problems, with learning taking place over the course of the exploratory process. A classic definition [18] of what makes ill-structured problems ill-structured is that the problem solver lacks sufficient knowledge to define the problem more precisely or enough knowledge to support search for a solution in a well-defined problem space. A particular problem may be ill-structured for a novice, but well-structured for a seasoned expert. In the context of information seeking, one might expect that people with domain knowledge would get less benefit from an exploratory search system (because their information-seeking in the domain will be more well-structured) than people with less domain knowledge (because their information seeking in the domain will be more ill-structured). More generally, a hypothesis is that, as users interact with exploratory search systems, they are supported in learning about particular domains.

Experimental Design

The experiment was a 2 (between-subjects) × 3 (within-subjects) mixed factorial design, with Interface (Exploratory vs. Baseline) as the between-groups factor, and subject matter domain (Future Architecture, Global Warming, and Web Mashups) as the within-subjects factor. Multiple tasks were performed to assess performance and learning.

METHOD

Participants

Thirty adults (22 male, 8 female) volunteered for this study from PARC (who received no compensation) or Stanford University (who were paid $40). Half were assigned to work with the full Exploratory MrTaggy condition and half worked with the Baseline condition. The participants'

average age was 31.9 years ranging from 21 to 54 years. Seventeen participants were native speakers of English; but the remaining thirteen also spoke English fluently. The majority of participants have either intermediate or advanced computer and Web search skills. They reported using computers (60 % of the participants over 35 hours a week) and the Web (50% of the participants over 25 hours a week) very frequently.

Interfaces

We compared the full, Exploratory MrTaggy interface (Figures 1 and 2) to a Baseline version of MrTaggy that only supported traditional query-based search (Figure 5). Both the Exploratory interface and the Baseline interface showed the search result snippets as presented in Figure 3. In both Exploratory and Baseline UIs, the snippets included presentation of a set of related tags. With both the Baseline and Exploratory UIs, users could directly type tags into the search box with a plus or minus sign as a prefix to reorder or filter a search results list. This method of query refinement was explicitly taught to users of both interfaces.

The Exploratory Interface additionally presented users with a related tags list down the left side of the UI with up and down arrows with which the user could provide relevance feedback (Figures 1 and 2). Clicking an up-arrow added the associated tag with a plus-prefix to the search box and invoked a reordering. Clicking a down-arrow added the associated tag with a minus-prefix to the search box and invoked a filtering. In other words, interaction with the related tags list in the Exploratory UI had the same effect as directly typing in tags (with plus/minus prefixes) into the search box. The Baseline UI did not include the related tags list or interactive arrows.

Task Domains

The experiment required participants to work through a series of information-seeking tasks in three different topic domains. The domains were selected to represent different
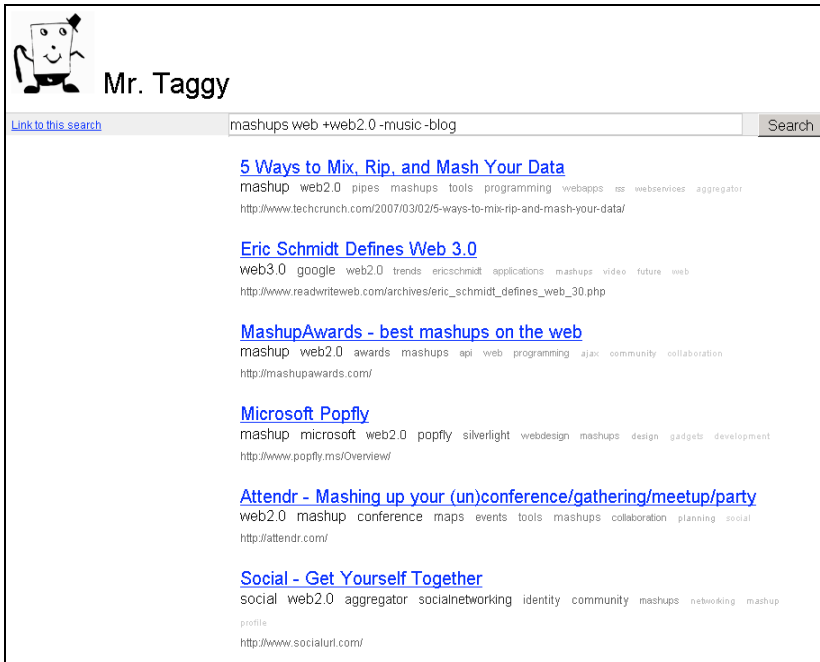
Figure 5. Baseline version of the MrTaggy user interface without related tags list on the left and without interactive relevance feedback. Users could still give feedback on tags by typing into the search box.

kinds of subject matter that might be encountered in everyday life. The domain of Future Architecture was selected as one exemplifying a creative domain; Global Warming exemplified a controversial domain; and Web Mashups a technology domain.

Moreover, the three topic domains differed in the level of ambiguity of the corresponding keywords [3]. In del.icio.us the tag "architecture" is highly ambiguous as this tag is often used for Web pages concerning software architecture as well as building architecture. The tag "mashups" is also partly ambiguous referring to both music and software. In contrast, "Global Warming" is less ambiguously tagged. In summary, we made some attempt to pick domains that varied across interesting dimensions.

Tasks

For each domain, prior to working with the MrTaggy interfaces, we assessed participants' prior knowledge with a battery of questions in a Prior Knowledge Test. During the main phases of the experiment, performance and learning was measured in three kinds of tasks: (1) finding results to pre-specified queries (Page Collection tasks), (2) writing (Summarization tasks), and (3) formulating keywords for search (Keywords formulation tasks).

Performance in the page collection task tested the effectiveness of the two interfaces in supporting the rapid collection of relevant pages—a task targeted by traditional (non-exploratory) search engines. The Summarization tasks and the Keywords tasks tested domain learning. We could also test whether the Exploratory UI (as compared to the

Baseline UI) compensated for a lack of prior domain knowledge in these learning tasks through tests for an interaction of Interface by Prior Knowledge on the Keywords and Summarization tasks.

For each domain, participants were asked to perform two Page Collection tasks, one Summarization, and one Keywords task. As described below, the Page Collection, Summarization, and Keyword tasks for each domain were done in sequence to foster any learning about the domain, prior to moving to the next domain, where the same tasks would be performed.

Prior Knowledge Test
At the beginning of the experiment, participants were asked to fill out a short computer-based questionnaire about their prior knowledge in the three topic domains. For each domain five to six general questions were presented to the participants, which all had to be rated on a 5-point scale (e.g., "How would you rate your knowledge about building Web 2.0 applications?" for the Web Mashups domain, "How would you rate your knowledge regarding environmental protection?" for the Global Warming domain, or "I could name a couple of architects or architecture firms spontaneously."). Cronbach's alpha was α =.92 for the Web Mashups scale, α =.84 for the Global Warming scale, and α =.65 for the Future Architecture scale. We did not use detailed questions by means of multiple-choice tests in order to avoid priming subsequent search processes.

Page Collection Task
In the two Page Collection tasks for each domain (see Table 1), participants were given a time limit and were requested to find as many pages as possible relevant to specific queries. The first Page Collection task was easy and the second difficult, based on the difficulty ratings obtained in pilot tests.

Domain	Task Difficulty	
	Easy	**Difficult**
Future Architecture	Pictures about Future Architecture	Architects or architecture firms from the US engaged in Future Architecture
Global Warming	Campaigns to fight Global Warming	Predictions about effects of Global Warming
Web Mashups	Examples of Web Mashups	How can Web Mashups be created

Table 1. Page Collection tasks 1 and 2 for the three topic domains.

In these tasks, the initial set of query words was predetermined, but participants could modify the query. In the Exploratory condition users could additionally provide relevance feedback as described before. Collecting a page was implemented by a "Save to collection" button.

Summarization Task
In the Summarization tasks (see Table 2) participants were given a time limit and asked to write a short coherent summary (max. 300 words) addressing one or two global questions or aspects concerning the topic domain. We hypothesized that, in contrast to the Page Collection tasks, Summarization required a more exploratory browsing strategy to acquire broader and more general conceptual understanding of the topic domain.

Participants were instructed to browse/search for the requested information and were restricted to include only information they found in their browsing. Users could move back and forth, and use cut-and-paste between the description page, Web, and summarization box and type or copy and paste the information into the summary box.

Domain	Summarization task
Future Architecture	Styles, forms and systems of architecture of the future: 1. Three different topics what "Future Architecture" could be about; 2. Summarize all 3 topics
Global Warming	Controversy about human-caused Global Warming: 1. Arguments or evidence in favor and against human-caused Global Warming; 2. Individuals/organizations who promote these arguments.
Web Mashups	Use of Web Mashups: 1. Benefits of the use of Web Mashups

Table 2. Summarization tasks for the three topic domains.

Keywords Task
In the Keywords tasks participants were given a time limit and were requested to generate and type in as many keywords as possible that were relevant to the corresponding topic domain.

Procedure
The experiment was conducted in a laboratory setting. Participants were provided an overview of the experiment and asked to fill out a short computer-based questionnaire to provide some demographic and personal data about their computer and internet usage and skills, as well as their prior knowledge concerning the three topic domains. Participants were then presented videos about the capabilities of the systems and the upcoming tasks (e.g., how to collect Web pages and how to type a summary into the "summary box", etc.). Participants were then assigned to work with either the Baseline or Exploratory MrTaggy system.

The participants then went through three blocks of tasks. Each block required the user to perform the Easy Page Collection, Difficult Page Collection, Summarization, and Keywords tasks, in that order, for one task domains.

The order of presentation of domain-blocks was counterbalanced across participants using a Latin Square. The Page Collection tasks were limited to 6 min each. The Summarization task was limited to 12 min. The Keywords task was limited to 2 minutes.

Between blocks, participants were asked to fill out a computer-based questionnaire to rate their subjective level of cognitive load during task processing using a modified version of the NASA task load index questionnaire [10].

Finally, subjects rated the use of the systems in a computer-based questionnaire. The whole experiment took around 2 hours.

RESULTS

Interaction Behaviors
To examine participants interaction behavior we analyzed: (1) the time taken, (2) the number of manually typed queries for query refinements, and (3) the number of overall queries, which in the Exploratory interface included participants' interactive relevance feedbacks. For each of these three variables we conducted a 2x9 MANOVA of Interface (Exploratory, Baseline) × Tasks (6 page collection tasks and 3 summary tasks).

For the number of overall queries, there was a main effect of Interface (F (1, 28) = 11.36, MSE = 96.85, $p < .01$). With the Exploratory condition, participants were more engaged in query refinements with $M = 7.81$ queries compared to the Baseline participants who averaged only $M = 3.77$ queries. In contrast, there was no main effect of Interface ($F < 1$) on the number of typed queries.

These results show that the Exploratory users did not substitute their manual query typing behavior by the use of the relevance feedback, but used the opportunity of the relevance feedback as an additional way of query refinements thereby resulting in more query refinements. Hence, we conclude that through the use of the Exploratory interface a more intense exploratory search process was conducted.

In addition, there was a main effect of Interface (F (1, 28) = 8.55, MSE = 10.31, $p < .01$) on the time taken for the tasks. Exploratory users on average took $M = 7.74$ min to work on their tasks, whereas Baseline users only took $M = 6.60$ min. There was an interaction of Interface by Tasks (F (8, 224) = 3.92, MSE = 2.43, $p < .01$). Bonferroni-adjusted post-hoc tests showed that Exploratory users worked significantly longer on the summary tasks ($ps < .05$ for all three domains), but not on the page collection tasks (for Future Architecture and Global Warming tasks, both $ps > .20$; for Web Mashups tasks marginally significant effects of $p = .09$ and $p = .10$).

In line with the findings concerning the query refinements, these longer task processing times for the summary tasks seem to confirm our expectations of a more intense exploratory search process with the use of the Exploratory interface.

Page Collection Task

A 2×3×2 mixed-factorial ANOVA of Interface (Exploratory, Baseline) × Domain (Future Architecture, Global Warming, Web Mashups) × Difficulty (easy, hard) was computed on the number of pages collected. There was no main effect of Interface ($F < 1$).

There was a main effect of topic Domain on number of pages collected (F (2, 56) = 4.87, MSE = 9.19, $p < .01$). Post-hoc tests revealed that the two extremes concerning the level of ambiguity differed significantly: For the Global Warming tasks (low ambiguity) significantly more pages ($p < .05$) were collected (M = 6.37) than for the Future Architecture tasks (high ambiguity) (M = 4.67). For the Web Mashups tasks, participants collected on average M = 5.78 pages. There was an interaction of Interface by Domain (F (2, 56) = 5.79, MSE = 9.19, $p < .001$). Bonferroni-adjusted post-hoc tests showed that Exploratory users collected more pages (M = 7.03) than the Baseline users in the Web mashup domain (M = 4.53), $p < .05$.

There was a main effect of task Difficulty on number of pages collected (F (1, 28) = 7.27, MSE = 12.71, $p < .05$). An average of M = 6.32 pages were collected on Easy tasks, vs. M = 4.89 for Difficult tasks.

Analyses of the relevance of the collected pages yielded a similar pattern. In addition to (1) *number* of pages collected, we also analyzed (2) *sum of the relevance values* of the pages collected, and (3) *mean relevance value* of the pages collected, which was computed as the sum of the relevance values divided by the number of pages collected. The relevance ratings for the pages were determined in a side study in which 20 people hired through Mechanical Turk [12] rated the collected Web pages on a 5-point Likert scale (5=highly relevant). For each collected Web page the mean relevance from all 20 relevance ratings was computed. Statistical analyses yielded the similar patterns for these two additional metrics.

Summarization

The *quality* of the summaries was measured based on predefined topic-specific criteria. Two raters familiar with the summary tasks and the predefined criteria rated each sentence written in the summaries. An inter-rater reliability computed on a 30% subsample of the summaries yielded a Cohen's kappa of 0.73 for "Future Architecture", kappa = 0.74 for "Global Warming", and kappa = 0.71 for "Web Mashups." One rater scored the remaining summaries.

Summaries were rated based on the quality of the answers according to the task description. The "Future Architecture" summaries were rated regarding the *number of reasonable topics* (0-3 points) they mentioned about what Future

Architecture could be about and the *overall quality of the topic descriptions* (0-2 points per topic). The "Global Warming" summaries were rated regarding the *number of arguments* they mentioned in favor and against human-caused global warming and regarding the *number of individuals or organizations* advancing these arguments they listed. The "Web Mashups" summaries were rated regarding the *number of benefits* of Web Mashups mentioned and the *overall quality of the benefit description* (0–5 points per benefit).

6 univariate ANCOVAs were computed using Prior Knowledge Test centered scores as covariates on each separate domain. In the domain of Future Architecture, with the Exploratory interface, participants' summaries included a significantly higher number of reasonable topics (M = 2.67) than with the Baseline interface (M = 1.80), F (1, 26) = 8.75, MSE = 0.76, $p < .05$. In the domain of Global Warming, users of the Exploratory interface included a significantly higher number of arguments (M = 3.27) in favor and against human-caused global warming than users of the Baseline interface (M = 1.67), F (1, 26) = 7.04, MSE = 2.67, $p < .05$.

Also, in the Web Mashups domain, in the Baseline interface, Prior Knowledge correlated positively with the number of benefits and with the quality of the descriptions (r = .46, p = .09 and r = .51, p = .05). In contrast, in the Exploratory interface there were no significant correlations with Prior Knowledge (r = -.18, ns and r = -.11, ns). This result suggests that prior knowledge tends to have an effect on the summaries generated in some domains with the Baseline interface, but this relation is attenuated in the Exploratory interface. This suggests that the Exploratory interface is compensating for differences in prior domain knowledge. The keyword task analysis below contains further evidences to this effect.

Keywords

For the Keywords tasks, we coded and tallied the *number of reasonable keywords* about each topic domain. We omitted the initial search keywords (e.g., "Future Architecture"). Singular and plural forms of a word were counted as one keyword.

Analyses of covariance with Prior Knowledge as a covariate revealed significantly more reasonable keywords generated by the Exploratory interface users over the Baseline users for "Future Architecture", $t(26)$ = 1.87, SE = 7.43, $p < .05$, and for "Web Mashups", $t(26)$ = 2.69, SE = 2.97, $p < .01$, but not for "Global Warming", $t(26)$ = 0.82, SE = 11.61, ns.

Inspection of the data suggested that the number of keywords generated was correlated with Prior Knowledge for the Baseline interface, but not for the Exploratory interface. Linear model analysis of the within-subjects relation between Prior Knowledge and (log transformed) keywords generated showed a mean slope of 0.06 for the Exploratory interface, which was not significantly greater

than zero, $t(14) = 0.96$, $p = .18$. However, the slope of relation between prior knowledge and (log transformed) keywords for the Baseline interface was 0.32, which was significant, $t(14) = 1.86$, $p < .05$. Furthermore, the difference between the slopes for the Exploratory and Baseline conditions was marginally different, $t(18) = 1.40$, $p = .09$.

These results suggests that Prior Knowledge tends to have an effect on the number of reasonable keyword generated with the Baseline interface, but this relation is attenuated in the Exploratory interface, This suggests that the Exploratory interface is compensating for differences in prior domain knowledge.

Cognitive Load
The cognitive load experienced by participants (ranging from 0=very low to 100=very high) was analysed by a 2-way ANOVA (Interface x topic Domain). The repeated-measures ANOVA showed a significant main effect of Interface on cognitive load (F (1, 28) = 5.06, MSE = 1063.68, $p < .05$). Participants operating the Exploratory interface had a significant higher cognitive load ($M = 67.18$) than Baseline participants ($M = 51.69$).

A possible explanation for the higher cognitive load caused by the Exploratory interface is the greater amount of cognitive processing during exploratory search due to the additionally presented related tags and the relevance feedback. Hence, the higher cognitive load is a hint that a deeper processing and consequently more intense learning and investigation activities took place during Summarization task processing. However, we also have to admit, that the higher cognitive load might have arisen from a higher level of frustration in the difficult page collection tasks (see 'page collection').

Subjective Ratings
At the end of the experiment participants rated the use of the systems. Participants were presented a set of statements (e.g. 'The system was easy to use'), and asked to rate on a five-point scale (5=highly agree).

For the statement 'The system gave me ideas about what else to search for', Exploratory participants' ratings ($M = 4.07$) were significantly higher (t (28) = 2.74, $SE = 0.29$, $p = .01$) than Baseline participants' ratings ($M = 3.27$).

Moreover, additional statements only presented to Exploratory system participants showed rather high agreements: with $M = 3.93$ for the statement 'The tags displayed in the related tags list were useful to refine my queries', $M = 4.07$ for the statement 'The related tags list provided some interesting additional aspects', $M = 3.87$ for the statement 'I think the related tags list contributed to the effectiveness of my search' and $M = 4.33$ for the statement 'It was easy to operate the up and down arrows to add or exclude tags or search results'.

Furthermore, there was a marginally significant difference between the Exploratory and the Baseline interface concerning the preferred use of the Tag search browser (χ^2 (2) = 5.79, $p = .06$). Participants were asked if they preferred to use MrTaggy either for fact finding ("to search for specific information") or for exploratory search ("to browse for information and interesting things") or for both purposes. With the Exploratory interface 73.3% of the N=15 participants indicated to prefer the system for exploratory search, whereas only 6.6% (one person) rated for the fact finding, and 20% would like to use it for both. In contrast, with the Baseline interface participants (N=14, as one participant did not answer this question) were indecisive about the preferred use of the system. 40% indicated to prefer the system for fact finding, 33.3% for exploratory search, and 20% rated for both.

Summary of Findings
In this study, we analyzed the interaction and UI design of the tag search browser called MrTaggy. The main aim of our study was to understand whether and how our Exploratory tag search browser is beneficial for domain learning.

We compared the full, Exploratory MrTaggy interface to a baseline version of MrTaggy that only supported traditional query-based search. We tested participants' performance in three different topic domains and three different task types. The results show:

(1) User interactions during the experimental tasks confirmed that Exploratory system users took advantage of the additional features provided by the system, i.e. they used the opportunity of relevance feedback, without giving up their usual manual query typing behavior. They also spent more time on task and appear to be more engaged in exploration than Baseline participants.

(2) Performance data in the page collection task showed no general advantage of the Exploratory system over the Baseline system regarding the rapid collection of relevant pages. A possible reason for the lack of effect might be that the top-ranked search results returned by the system based on the given keywords were among the pages with highest rated relevance values. Even so, at least for the medium ambiguous Web mashup domain, Exploratory users did collect more pages with a higher sum of relevances than the Baseline users.

(3) For learning outcomes our expectations were partly confirmed as there are some indications for summaries of higher quality with the Exploratory system compared to the Baseline system. More precisely, Exploratory system users' summaries included a higher number of reasonable topics about Future Architecture, a higher number of arguments in favor and against human-caused global warming.

(4) Also to gauge learning outcomes, with respect to the Keyword Tasks, Exploratory system users generated more reasonable keywords than the Baseline users for the two topic domains of medium and high ambiguity "Web

Mashups" and "Future Architecture", but not for the low ambiguity domain "Global Warming".

(5) The Exploratory UI compensated for (i.e., attenuated) effects due to differences in prior knowledge in one of the three Summarization tasks and two out of three Keywords tasks.

DISCUSSION
The results above suggest that the exploratory functions of the tag search browser appear to be beneficial for domain learning. Results show that subjects with the Exploratory interface are more engaged with their tasks. One indication of higher engagement was that people using the Exploratory interface spent more time writing the summaries. A second indication of higher engagement was that people using the Exploratory interface reported higher cognitive load. Through the use of the Exploratory interface subjects conducted a more intense exploratory search process.

Interestingly, there are some indications that the Exploratory tag search system is particularly beneficial for partly ambiguous keywords. The different meanings of a word might not come to mind spontaneously so that the presentation of related tags can support the users in their query refinements. As evidence, in the medium ambiguous Web mashup domain, Exploratory users collected more pages with a higher sum of the relevance values than the Baseline users.

More importantly, the results of the Summarization tasks and Keyword tasks at least partly confirmed our hypothesis that users interacting with the Exploratory interface are supported in their learning and investigation activities. Exploratory interface users wrote summaries that contained more detail in two out of the three Summarization tasks and generated more reasonable domain keywords in two out of three Keywords tasks. These indicate higher domain learning outcomes compared to a search system without related tags and little support for interactive relevance feedback.

Moreover, results from analyzing prior knowledge in the Summarization and Keyword tasks suggest that the Exploratory tag search system is particularly beneficial for novice users of a topic area to gain domain knowledge. The full Exploratory interface seems to offer a kind of scaffolding support for novice users to perform as well as expert users, enabling participants to perform at a high level, regardless of their level of prior domain knowledge.

In summary, the results of the study indicate a particular benefit of our Exploratory tag search system in supporting users in their exploratory search in order to gain new knowledge in ill-structured domains. This conclusion is further strengthened by the high percentage (73.3%) of Exploratory system users' subjective preference to use MrTaggy for exploratory search. Thus, the functionality of our Exploratory tag search system is promising and we plan to continue our work in order to further improve the system and to strengthen and generalize the results of this study.

Limitations
There are some obvious limitations to our study.

First, we were limited in the choice of our subjects. Prior domain knowledge of subjects was only measured by a short and rather general domain knowledge questionnaire. Therefore, future research is needed to explicitly compare performance of pre-selected domain novices, semi-experts, and experts when interacting with either the Exploratory or the Baseline system. Furthermore, the sample size should be increased in order to increase statistical power.

Second, the levels of complexity and ambiguity of the three topic domains were defined by the experimenters prior to the study, but have not been validated by external ratings and more objective measures. Thus, in future work a broader range of topic domains with clearly defined levels of ambiguity will be used to receive more detailed insights in the relationship between topic ambiguity and benefit of the Exploratory tag search system.

Third, we tested both the interactive relevance feedback feature by means of up and down arrows and the presentation of the related tags list integrally. Thus, it cannot be differentiated, whether the advantage of our Exploratory tags search system is due to the related tags list presented, or due to the interactive relevance feedback feature or due to the combination of both features. Hence, in future experiments a third condition should be included presenting the related tags list without relevance feedback. Furthermore, a fourth even more Baseline condition could be investigated which neither provides relevance feedback and the related tags list, nor presents the related tags in the search result snippets.

Fourth, in order to reduce subject variability, our experimental procedures included both starting query words as well as some interface training for query refinement. For each topic domain, an initial set of query words was predetermined. Hence, this might have unnaturally unified users' search behaviors. The predetermined query words might have induced a rather passive behavior and thus might have hindered users in applying their own personal search strategies, deeper processing, and creative thinking. Moreover, detailed experimental instructions also included an explanation of query refinement capabilities of both interface conditions, which might have strongly increased their application.

Finally, by predefining the topic domains the study did not address any personal information needs, which might have also lead to higher engagement with the search process.

To increase ecological validity, future experiments might exclude some training instructions and allow subjects to search for subjects of their own interest. While this would decrease the power of the experiment, but we would then be able to test subjects in a more naturalistic setting.

CONCLUSION

This paper has introduced MrTaggy, an exploratory tag search browser that allows users to explore social tagging data in order to learn about unfamiliar domains. We dealt with tag noise algorithmically by computing tag and URL co-occurrence patterns. The empirical results show that subjects can effectively use data generated by social tagging as "navigational advice".

The study's first insights regarding the use of our exploratory tag search system are promising that the tag search browser can support users in their exploratory search process. The results suggest that users' learning and investigation activities are fostered by both relevance feedback mechanisms as well as related tag ontologies that give a kind of scaffolding support to domain understanding. Although further research is needed, the experimental results provide first indications that users' explorations in unfamiliar topic areas can be supported by the domain keyword recommendations presented in the related tags list and the opportunity for relevance feedback provided by the system.

Finally, since search engines that depend on social cues rely on data quality and increasing coverage of the explorable Web space, we expect that the constantly increasing popularity of social bookmarking services among different kind of users will improve social search browsers like MrTaggy. The results of this project point to the promise of social search engines and browsers to fulfill a need in providing navigational signposts to the best contents out in the vast Web.

REFERENCES

1. Baeza-Yates, R. and Ribeiro-Neto, B. *Modern Information Retrieval.* Addison Wesley, 1999.

2. Bush, V. As We May Think. *The Atlantic Monthly*, *176,* 1, (1945),101-108.

3. Chi, E.H. and Mytkowicz, T. Understanding the Efficiency of Social Tagging Systems using Information Theory. *Proc. Hypertext 2008*, ACM Press (2008),. 81-88.

4. Evans, B. and Chi, E. H. Towards a Model of Understanding Social Search. In *Proc. of Computer-Supported Cooperative Work (CSCW)*, pp. 485-494. ACM Press, 2008. San Diego, CA.

5. Fitzpatrick, L. and Dent, M.. Automatic feedback using past queries: social searching? *Proc. 20th Annual Intern. ACM SIGIR Conference.* ACM Press (1997), 306-313.

6. Furnas, G.W., Fake, C., von Ahn, L., Schachter, J., Golder, S., Fox, K., Davis, M., Marlow, C. Naaman, M.. Why do tagging systems work? *Extended Abstracts CHI 2006*, ACM Press (2006), 36-39.

7. Furnas, G.W., Landauer, T.K., Gomez, L.M. and Dumais, S.T.. The vocabulary problem in human-system communication. *Communications of the ACM*, *30* (1987), 964-971.

8. Glance, N.S. Community search assistant. *Proc. IUI 2001*, ACM Press (2001), 91-96.

9. Golder, S. and Huberman, B.A. Usage Patterns of Collaborative Tagging Systems. *Journal of Information Science*, *32,* 2 (2006), 198-208.

10. Hart, S.G. and Staveland, L.E. Development of the NASA-tlx (Task Load Index): Results of empirical and theoretical research. *Human Mental Workload* (1988), 139-183.

11. Hong, L., Chi, E.H., Budiu, R., Pirolli, P. and Nelson, L. SparTag.us: Low Cost Tagging System for Foraging of Web Content. *Proc. AVI 2008*, ACM Press (2008), 65-72.

12. Kittur, A., Chi, E.H., and Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proc. CHI 2008,* ACM Pres (2008), 453-456.

13. Lee, K.J. What Goes Around Comes Around: An Analysis of del.icio.us as Social Space. *Proc. CSCW'06*, ACM Press (2006), 191-194.

14. Marchionini, G. Exploratory search: From finding to understanding. *Communications of the ACM*, *49, 4* (2006), 41-46.

15. Millen, D., Yang, M., Whittaker, S.and Feinberg, J. Social bookmarking and exploratory search. In L. Bannon, I. Wagner, C. Gutwin, R. Harper, and K. Schmidt (eds.). *Proc. ECSCW'07,* Springer (2007) 21-40.

16. Raghavan, V.V. and Sever, H. On the Reuse of Past Optimal Queries. *Proc. SIGIR95*, ACM Press (1995), 344-350.

17. Shneiderman, B., Byrd, D. and Croft, W.B. Clarifying search: a user-interface framework for text searches, *D-lib magazine*, *3,* 1 (1997), Available at http://www.dlib.org/dlib/january97/retrieval/01shneiderman.html.

18. Simon, H.A. Structure of ill structured problems. *Artificial Intelligence*, *4*, 3-4 (1973), 181-201.

19. White, R.W., Drucker, S.M., Marchionini, M., Hearst, M., schraefel, m.c. Exploratory search and HCI: designing and evaluating interfaces to support exploratory search interaction. *Extended Abstracts CHI 2007*, ACM Press (2007), 2877-2880.