Aslib Journal of Information Management

Two ' s company, but three ' s no crowd: Evaluating exploratory web search for individuals and teams

Chirag Shah Chathra Hendahewa Roberto González-Ibáñez

Received 13 May 2015
Revised 14 September 2015
Accepted 21 September 2015

Two's company, but three's no crowd

Evaluating exploratory web search for individuals and teams

Chirag Shah and Chathra Hendahewa
Rutgers University, New Brunswick, New Jersey, USA, and
Roberto González-Ibáñez
Universidad de Santiago de Chile, Santiago, Chile

Abstract

Purpose – The purpose of this paper is to investigate when and how people working in collaboration could be benefitted by an exploratory search task, specifically focussing on team size and its effect on the outcomes of such a task.

Design/methodology/approach – The paper investigates the effects of team sizes on exploratory search tasks using a lab study involving 68 participants – 12 individuals, ten dyads, and 12 triads. In order to assess various factors during their exploratory search sessions, an evaluation framework is synthesized using relevant literature. The framework consists of measures for five groups of quantities relevant to exploratory search: information exposure, information relevancy, information search, performance, and learning.

Findings – The analyses on the user study data using the proposed framework reveals that while individuals working alone cover more information than those working in teams, the teams (dyads and triads) are able to achieve better information coverage and search performance due to their collaborative strategies. In many of the measures, the triads are found to be even better than the dyads, demonstrating the value of adding a collaborator to a search process with multiple facets.

Originality/value – The findings shed light on not only how collaborative work could help in achieving better results in exploratory search, but also how team sizes affect specific aspects – information exposure, information relevancy, information search, performance, and learning – of exploratory search. This has implications for system designers, information managers, and educators.

Keywords Evaluation, Group work, User study, Information seeking, Collaborative search, Exploratory search

Paper type Research paper

1. Introduction

Collaboration is considered a useful trait in many situations (Denning, 2007; Denning and Yaholkovsky, 2008). These situations often involve working with information. Examples include patent searching (Hansen and Järvelin, 2005), knowledge work (Morris, 2008), healthcare (Reddy and Jansen, 2008), and education (Hyldegård, 2006). However, our knowledge about when and how people working in collaboration could be benefitted in an exploratory search task is limited. Researchers have shown the effectiveness of pairs of searchers working on an information retrieval (IR) task (e.g. Pickens *et al.*, 2008), but these works have two major limitations: they are either employing traditional IR measures such as precision and recall to evaluate goodness of collaborative search, or too limited in their scope. The latter includes the kind of tasks and team sizes (typically two). Some have experimented with exploratory search tasks (e.g. Shah and González-Ibáñez, 2011), but still staying focussed on the searching aspects of exploratory search. Others have studied more than two people working

Aslib Journal of Information Management
Vol. 67 No. 6, 2015
pp. 636-662

2050-3806
DOI 10.1108/AJIM-05-2015-0082

together (e.g. Morris *et al.*, 2006), but lacked comparative conditions to measure the effects of group size in a collaborative search task. In the work reported here, these limitations are addressed by considering an exploratory search task, and experimenting with different group sizes: 1 (individuals), 2 (dyads), and 3 (triads).

Specifically, the current paper adopts a user-centric perspective to exploratory search evaluation by investigating what different kinds of quantitative metrics captured from the log data of users conducting exploratory search tasks could help us in evaluation of search. Further, this paper evaluates not only individual users but also users working collaboratively on the same exploratory search task. Thus, the paper has two primary contributions: first, presenting a framework by synthesizing and extending existing measures for evaluating exploratory Web search; and second, applying this framework to evaluating searches performed by individuals, dyads, and triads. The former portion is neither a unique contribution nor the focus of this paper; it is based on previous works and necessary here for the experiments. The latter portion is demonstrated by analyzing data collected through a lab study with 68 participants.

The remainder of the paper is organized as follows. An overview of evaluation measures in IR is presented in the next section, focussing specifically on exploratory and collaborative searches. A framework for evaluating exploratory search using Web search and related measures categorized into five groups of analysis is then presented in the evaluation framework section. A user study is described in the section titled user study, which includes a description of the exploratory search task, participants, and the data logged. The results of this study, which are obtained using the synthesized evaluation framework, are presented in the results section along with interpretations. This is then followed by discussions and limitations in the discussion section. The paper finishes in the conclusion section with the findings generated from this approach and a discussion of what this work adds to the evaluation of exploratory Web search in general and collaborative search in particular.

2. Background

This section provides a brief background on two areas: exploratory search and collaborative search. In both cases, the concepts are introduced, followed by a discussion regarding the forms of measurement used to evaluate the concept. Specifically, this section will suggest the difficulties in evaluating exploratory Web search in general and collaborative search in particular by listing some of the efforts for evaluation in the literature and identifying their shortcomings.

2.1 Exploratory search

Exploratory search can be characterized by five major factors: uncertainty, ambiguity, low level of specificity, and the need for discovery and learning from multiple sources (Marchionini, 2006; Kules and Capra, 2009; Wildemuth and Freund, 2012). White and Roth (2009) define exploratory search as "a type of information seeking and a type of sense-making focused on the gathering and use of information to foster intellectual development." Unlike other types of search tasks such as fact-finding, exploratory search is open-ended (Wildemuth and Freund, 2012). This is partially due to the dynamic and evolving nature of this type of task. In addition, it is often acknowledged that designing exploratory search tasks to conduct research can be especially challenging (Kules and Capra, 2009; Wildemuth and Freund, 2012). Yet this type of task is widely used in information-related studies.

White and Roth (2009) introduced exploratory search as follows: "what if we want to find something from a domain where we have a general interest but not specific knowledge?" (p. 37). For instance, global warming is a recurrent topic that has grabbed the attention of many, however, knowledge on this subject for lay people may be limited to shallow aspects. Even when clear goals are specified (e.g. causes and consequences), research on global warming would require exploration and evaluation of multiple sources. Yet this may not be sufficient to expose searchers to the different views on this matter. The complexity of exploratory search is particularly clear when current technology mediates access to information. For example, how are information needs translated into queries? To what extent do such queries lead to meaningful information? When does the information gathered fulfill the original information need? And finally, for researchers, evaluating exploratory search has always been a big research question.

According to Broder (2002) as well as Rose and Levinson (2004), Web search queries can be generally categorized in three major classed based on the searcher goals and intents namely, informational, navigational, and resource or transactional. In the context of exploratory search, the type of queries executed by searchers are mostly informational in nature that enable searchers to address an information need. Broder (2002) defines the intent behind informational queries as "to acquire some information assumed to be present on one or more web pages in their static form and static form means the target document is not created in response to the user query" (p. 5). Unlike informational queries, navigational and transactional queries are more specific in which the user's goal is to find a Web site or to perform some Web-mediated activity that is not highly visible in exploratory search. Jansen *et al.* (2008) operationalized a taxonomy that can automatically categorize user search queries into informational, navigational, or transactional types. With respect to their three-level hierarchical taxonomy used to classify the existing user-intent-based studies, they show that exploratory search would have a variety of informational type queries ranging from quickies (quick fact queries for advice) to informational queries that are used to gather, collect, and locate information. Their analysis showed that more than 80 percent of Web queries are informational in nature thus, showing the importance of analyzing informational queries that could be part of exploratory search described in this paper. Part of the taxonomy for information queries consists of question words, natural language terms, information terms, words beyond the first query, query length greater than two which are also present in most of the queries associated with exploratory search.

Evaluation is one of the core issues in IR. Numerous measurements have been developed to evaluate various aspects of user and system search performance (Saracevic, 1995). These include traditional measures such as precision, recall, mean average precision (MAP), and normalized discounted cumulative gain (NDCG), as well as less used measures such as novelty and mean reciprocal rank (MRR). The usage of any of these measures is driven by the context and application. For instance, recall is useful for patent finding and MRR is more relevant for homepage finding or high-accuracy (Allan, 2004) tasks.

When it comes to exploratory Web search, it becomes less clear which measures should be used. As the name implies, exploratory search is the type of searching that involves the information seekers undergoing complex cognitive tasks that lead to learning, exploration, and acquiring intellectual skills. Some of the main characteristics of exploratory search that differentiate it from other types of searches are uncertainty,

creativity (Bawden, 1986), innovation, knowledge discovery, serendipity (Foster and Ford, 2003), learning, and investigation (White and Roth, 2009, p. 10; Marchionini, 2006; Budd, 2004). White and Roth (2009) explained that due to the complexity and the uniqueness of exploratory search, evaluating this process using traditional IR measures based on retrieval accuracy, such as precision, recall, MRR, MAP, and NDCG, may be inappropriate. Therefore, it is important to investigate measures and frameworks that could ultimately capture the major characteristics integral to exploratory search, focussing on user-interaction and user behavior.

Two possible ways to evaluate exploratory search are system-centric and user-centric. Past research has mainly focussed on the development of new systems and user interfaces that support exploratory search (White and Roth, 2009, p. 61). This has led to the development of exploratory search evaluation metrics using a system-centric approach. Some of the metrics identified as candidate measures in evaluating the performance of exploratory search systems (White *et al.*, 2006a) are engagement and enjoyment, information novelty, task success, task time, learning, and cognition.

Taking an alternative perspective to evaluating exploratory search, one could evaluate the search performance taking a user-centric view where characteristics of the actions and behaviors of users conducting exploratory search tasks can be studied. Few such attempts have been made to evaluate exploratory search performance by applying the user-centered approach developed by Spink (2002) with the use of a new search tool covering aspects of effectiveness and usability. The analysis was based on relevance judgments, users' pre and post search questionnaires, and search logs (qualitative analysis of audio taped sessions). To the best of our knowledge, there have not been extensive attempts to identify metrics to evaluate exploratory search performance from a user-centric point of view, other than using qualitative measures using surveys and interviews.

Whether exploratory search is performed individually or collaboratively, evaluation is described as one of the major challenges (Kules and Capra, 2009; Wildemuth and Freund, 2012). This includes specific subtopics such as the evaluation of systems that support exploratory search (White *et al.*, 2006a, b) and also the evaluation of search performance (White *et al.*, 2006a, b). In this regard, how should exploratory search tasks be designed to perform empirical evaluations? In a given exploratory search task, how can search products from different searchers be compared and contrasted with each other? Are there implicit indicators that indicate the satisfaction of searchers? In the context of collaborative exploratory search, it is of interest to determine if the associated costs of solving this type of search task collaboratively with others are compensated by the potential benefits. To address this problem, evaluation frameworks involving traditional (e.g. precision and recall) and non-traditional measures (e.g. likelihood of discovery (LD) and unique coverage) have been proposed (Shah and González-Ibáñez, 2011). Yet in most cases, studies have been limited to individual searchers and pairs. This paper advances the previous efforts in evaluation frameworks by covering the constituent parts of exploratory search using metrics calculated from a user-point of view and extending these not only to individual searchers and pairs but, also to groups of three. The approach proposed in this paper provides a general framework that facilitates the evaluation of exploratory search for individual searchers and teams of searchers working collaboratively. This framework highlights aspects such as information search, information relevancy, information seeking, search performance, and learning that are explained in detail in the next section.

2.2 Collaborative search

For many years, IR has focussed on individual users searching for information (Foster, 2006). Algorithms have assumed that one person is reviewing the results and user interfaces have supported the needs of individual searchers. Yet, social behaviors that surround processes such as information seeking and searching have been recognized as part of information behaviors. For instance, Wilson's (1981) model of information behavior describes information exchange as a collaborative conduct that derives from information seeking. In this case, information exchange refers to the process whereby people provide and/or obtain perceived-useful information to and from others. Allen's (1997) integrated "person-in-situation-behavior" model distinguishes individual and collective information needs. Likewise, Sonnenwald (1999), Sonnenwald and Pierce (2000), and Talja (2002) focussed on information needs of groups highlighting the participation of social factors as part of information behaviors.

Collaborative search focusses on the notion that information seeking is not always a solitary activity and that people working in collaboration in information seeking tasks should be studied and supported. Collaborative search is also referred to as collaborative exploratory search (Pickens and Golovchinsky, 2007), collaborative information seeking (Shah and González-Ibáñez, 2011; Foster, 2006), and collaborative IR (Fidel *et al.*, 2004). Recently, there has been a surge of interest in this topic from both industry and academia surrounding issues of awareness (Shah and Marchionini, 2010), algorithmic mediation (Pickens *et al.*, 2008), time/space (González-Ibáñez *et al.*, 2013), roles (Shah *et al.*, 2010; Soulier *et al.*, 2014), and other collaborative aspects to develop, evaluate, and deploy software tools and algorithms that support collaborative search[1].

While most research on exploratory search has focussed on individual searchers, there are also studies that focus on collaborative practices around this type of search task (Pickens and Golovchinsky, 2007). In such cases the information needs that trigger exploratory search behaviors are common for two or more individuals who collaborate in different temporal and spatial settings (Pickens and Golovchinsky, 2007; González-Ibáñez *et al.*, 2012). Due to the complexity of exploratory search tasks in collaborative settings, a common behavior described in the literature is division of labor, which depending on the topic and structure of the task, allows group members to divide up the work (Morris, 2007, 2008).

Evaluating a collaborative search environment can be a huge challenge due to its complex design that involves a set of users, integrated systems, and a variety of interactions. One can evaluate a collaborative search system using typical measures of IR. However, information seeking is not merely about retrieving information, and thus, evaluating a collaborative search system by its retrieval effectiveness may not be sufficient. While traditional IR evaluations can still be used to measure the retrieval performance of a collaborative filtering system, just as Smyth *et al.* (2005) did, one needs additional measures for collaborative search systems.

Baeza-Yates and Pino (1997) first presented some initial work on developing a measure that can extend the evaluation of a single-user IR system for a collaborative environment. While this was based on retrieval performance, Aneiros and Estivill-Castro (2005) proposed to evaluate the goodness of a collaborative system by its usability. In addition, Baeza-Yates and Pino (1997) treated the performance of a group as the summation of the performances of individuals in the group. While this may work for simple information seeking and retrieval tasks, one can imagine situations in which this is not true. For instance, if two people working together can find twice

as much information as either of them working independently, was that a good thing? What about the amount of time they spent cumulatively? The participants may not be able to find twice as many results, but what if they achieved a better understanding of the problem or the information due to working collaboratively? Then there are other factors, such as engagement, social interactions, and social capital, which may be important depending upon the application, but are usually not looked at in non-interactive or single-user IR evaluations.

Wilson and schraefel (2008) analyzed an evaluation framework for information seeking interfaces in terms of its applicability to collaborative search software. Extending Bates' (1979) tactics model and Belkin *et al.*'s (1993) model of users, they showed that the framework could be just as easily applied to collaborative search interactions as individual information seeking software. But they also pointed out that there are additional considerations about the individual's involvement within a group that must be maintained as the assessment is carried out.

To overcome the issues with existing measurements and frameworks for evaluating individual and collaborative exploratory search, a new framework is proposed in the following section. This framework is essentially a synthesis of relevant methods and measurements for capturing different aspects of (individual) exploratory search.

3. Evaluation framework

It is clear from the literature that in an exploratory search session, an information seeker starts with a vague notion of the topic and the information need and moves toward a better understanding of both. In this paper, an attempt has been made to incorporate different aspects of exploratory search that are considered to be relevant in evaluating its open-ended, multi-faceted, dynamic nature. Two such aspects are information coverage (refers to information exposure), and situational relevance (refers to information relevance) as described by Saracevic (2007). Further, it is described in White *et al.* (2006a, b) that "[…] users generally combine querying and browsing strategies to foster learning and investigation" (p. 38). This claim shows that analyzing the query execution and clicks on search engine results pages (SERP) is essential in evaluating the search behavior of users conducting exploratory search thus, that aspect is incorporated under the heading of information search. Another important aspect to measure is the performance level of users in finding information to satisfy the information need to evaluate whether the user is heading in the correct direction in acquiring the information. Final aspect that is considered here is learning, which is viewed to be a major constituent of differentiating other type of information searches from exploratory search, that involves complex cognitive loads and adding knowledge about the topic of interest over the search process (White and Roth, 2009; Marchionini, 2006).

Therefore, this evaluation framework attempts to capture the variety of ways users try to express the information need and rectify it as needed (information search) and how much reasonable amount of information they are exposed to (information exposure) while searching. As explained above, a successful completion of an exploratory search task will also include exploring and collecting relevant information (information relevancy) with high effectiveness and/or efficiency (performance). Finally, as Marchionini (2006) and White and Roth (2009) have pointed out, learning is an important aspect within exploratory search; and therefore, it is incorporated to the framework to make the evaluation framework comprehensive by spanning across five different aspects (mentioned above) of exploratory search

that can be evaluated using log data analysis. Fortunately, there are some recent works in the literature that could be used for synthesizing an evaluation framework that would work here. And therefore, once again it is pointed out that many of the measures presented here are based on Shah's (2014b) recent article. Specifically, most of the measurements for information exposure, information relevancy, information seeking, and performance are either taken directly or in a modified form from the aforementioned article. Most of the measures presented below have been previously used (Shah and González-Ibáñez, 2011, 2012) for individuals and pairs of searchers, but not for triads. Note that the measures described here are designed primarily for evaluating exploratory search for individuals. However, the objective here is to apply them to collaborative settings not only to show how they could be successfully leveraged to measure collaborative IR, but also to compare individual and collaborative IR in exploratory search tasks using the same measures. Also note that throughout the paper, individual user or a team is denoted as t and all users or all teams as T.

3.1 Information exposure

This refers to the amount of information one discovers through active searching or passive browsing while working on an exploratory search task. In the case of the Web, such information primarily refers to the visited Web pages. A typical IR measure for evaluating information exposure or retrieval is recall, but one could also compute coverage of information as given:

$$Coverage_t = \{wp_1, wp_2, \ldots\ldots, wp_n\} \tag{1}$$

In (1), wp_i denotes distinct Web pages visited by user/team t.

The universe of distinct Web pages and the universe of relevant pages visited by all user/teams, $\forall t \in T$ are defined as U and U_r, respectively:

$$U = \cup_t \, Coverage_t \tag{2}$$

$$U_r = \cup_t \, RelevantCoverage_t \tag{3}$$

In (3), RelevantCoverage$_t$ is the set of Web pages that user/team t visited and found as relevant by collecting snippets from. This assumption, i.e., the union of relevant pages visited by the participants is the whole universe of relevant pages, while not completely supported, is a reasonable method to provide a quantifiable ways to measure quantities such as recall and precision as shown by (Shah and González-Ibáñez, 2011, 2012; González-Ibáñez *et al.*, 2012). Another measure of coverage consists of all Web pages within the coverage of a given user/team t that were visited only by t and not by any other user/team in the set of users/teams T. This measure, referred to as UniqueCoverage$_t$, is expressed in the following equation:

$$UniqueCoverage_t = Coverage_t \notin \cup_{t_i \in T \notin \{t\}} Coverage_{t_i} \tag{4}$$

The traditional IR measure of recall in this context is computed as follows:

$$Recall_t = \frac{|RelevantCoverage_t|}{|U_r|} \tag{5}$$

In addition, it may be useful to understand how difficult it was to discover the information that one did. Shah and González-Ibáñez (2011, 2012) referred to this as LD, and it is defined in the following equation:

$$LD_{wp_i} = \frac{-1 \cdot n\{wp_i\}}{|U|} \tag{6}$$

LD_{wpi} is the LD value for each Web page, wp_i. $n\{wp_i\}$ refers to the frequency of Web page wp_i appearing in the entire collection of Web pages denoted by Coverage_t in Equation (1). Based on the above measure, LD value can be found for all Web pages visited by each user/team t as defined below. This measure gives a lower value to Web pages that were visited by many users/teams within the corpus while resulting in a higher value to Web pages that were only visited by few users/teams. Thus, this measure tries to capture how easy or difficult it was to find information under the assumption that if many users/teams found that piece of information, then it was easy to find whereas if only a few users found that particular information then it was more difficult to discover.

Once LD_{wpi} is calculated for each Web page, those measures are summed over the set of Web pages visited by each user/team and divide by the number of coverage corresponding to that specific user/team in arriving at the final LD score for each user/team t, as denoted in the following equation:

$$LD_t = \frac{\sum_{i=1}^{|Coverage_t|} LD_{wp_i}}{|Coverage_t|} \tag{7}$$

3.2 Information relevancy

This refers to the relevance of information discovered to satisfying the information need of the search task. In IR, this is typically measured by precision or an extension/variation of it such as MAP that are considered to be not appropriate for exploratory search evaluation. Therefore, an alternative approach is taken by looking at the relevancy of the covered information by observing the snipping behavior where users collected snippets from Web pages to be used in their final reports.

Measures based on the intersection of coverage, relevance, and uniqueness are described in the following equations that are used to measure the information relevancy criteria of each user/team, t:

$$RelevantCoverage_t = Coverage_t \cap U_r \tag{8}$$

$$UniqueRelevantCoverage_t = UniqueCoverage_t \cap U_r \tag{9}$$

$$NumSaved_t = |\cup_t SnippetsCollected_t| \tag{10}$$

The traditional IR measure of precision in this context is computed as in the following equation:

$$Precision_t = \frac{|RelevantCoverage_t|}{|Coverage_t|} \tag{11}$$

3.3 Information search

Information search refers to "the behavioral manifestation of humans engaged in information seeking and also to describe actions taken by computers to match and display information objects" (Marchionini, 1995, p. 5). Here, this refers to the way information is sought and retrieved. In the case of Web search, the most common way of searching for information is by submitting queries to Web search engines. Therefore, measures that look at quality and quantity of queries should be considered for this aspect of exploratory search.

The set of distinct queries issued by user/team t and the set of SERP clicked by each user/team t were measured as shown in the following equations, respectively:

$$Q_t = DistinctQueries_t \tag{12}$$

$$SERP_t = SERPclicked_t \tag{13}$$

Levenshtein distance (based on string characters) between each pair of query strings, Q_a and Q_b within the set of distinct queries executed by each user/team t are found and averaged. Those distances shown in the following equation, measure the level of difference/diversity between the queries each user/team issued:

$$QueryDiversity_t = mean\{LevenshteinDistance\{Q_a, Q_b\}\}, Q_a \neq Q_b \wedge \{Q_a, Q_b\} \in Q_t \tag{14}$$

In order to define the information content of each query string Q_a the information entropy measure can be used as defined:

$$Entropy_{Q_a} = \sum_{u=1}^{|unigrams_{Q_a}|} -p_u \log_2 p_u \tag{15}$$

In Equation (15), p_u is the frequency of counts of each unigram, u appearing in each query string, Q_a found in the entire data set. The information content for each user/team t can be found as the mean of entropy values of each distinct query issued by user/team t as shown in the following equation:

$$AvgInfoContent_t = \frac{\sum_{a=1}^{NQ_t} Entropy_{Q_a}}{|Q_t|} \tag{16}$$

3.4 Performance

This refers to overall goodness of the search process. A typical measure in IR is F-score, which combines recall (information exposure) and precision (information relevancy):

$$F_t = \frac{2 \cdot Precision_t \cdot Recall_t}{Precision_t + Recall_t} \tag{17}$$

To understand how quickly and effectively a user is finding useful information, a measure relating to dwell time is employed. Specifically, the number of Web pages where the dwell time (time spent on the Web page) is more than 30 seconds can be considered as a useful Web page (Fox *et al.*, 2005; White and Huang, 2010).

Using this measure of usefulness of a Web page, effectiveness and efficiency (González-Ibáñez *et al.*, 2012) for each user/team t can then be defined as shown in the following equations, respectively. Here, effectiveness measures how much of the

covered information landscape was useful, whereas efficiency measures the amount of effectiveness achieved per search (query):

$$Effectiveness_t = \frac{\left|\cup_t \{wp_i(DwellTime_{wp_i} \geqslant 30secs)\}\right|}{|Coverage_t|} \tag{18}$$

$$Efficiency_t = \frac{Effectiveness_t}{|Q_t|} \tag{19}$$

3.5 Learning

An important aspect of exploratory search is the learning that takes place in the searcher as he proceeds with the task. A common way to measure this is by using pre-task and post-task questionnaires that reveal how much the searcher knows about the topic before doing the task as well as the level of confidence gained in their findings after performing the task. However, since the objective here is to evaluate different quantities by analyzing log data, a different approach is needed that includes tracking changes in other quantities over time.

To understand learning through analyzing log data, one needs to see how different factors or behaviors about search change with time as the users are performing their task. If, on the other hand, no learning took place, one can form the following null hypotheses using the other four forms of measures described earlier:

- $H_0 1$. A user/team will continue finding information with equal difficulty to discovery (information exposure).
- $H_0 2$. A user's/team's coverage of relevant information will not increase with time (information relevancy).
- $H_0 3$. A user's/team's expression of information need will not vary with time (information search).
- $H_0 4$. A user's/team's performance will remain unchanged with time (performance).

To test these hypotheses, relevant measures are needed for both individual user-level and team-level data for every minute of the session to create a time-series. While a measure for evaluating performance exists in the proposed framework (effectiveness), it is not suitable for seeing how well a user/team t does with respect to time. This is because the universe of information changes with time. It will eventually become harder to find useful information or maintain the same level of effectiveness as measured by Equation (18). A modified measure is therefore suggested that takes into consideration the remaining universe of unexplored information at a given time. It is called effectiveness of discovery (EoD), defined for each user/team t for each time point k as shown in the following equation:

$$EoD_{t,k} = \frac{\left|\cup_{t,k} \{wp_i(DwellTime_{wp_i} \geqslant 30secs)\}_{t,k}\right|}{\left|U - Coverage_{t,k}\right|} \tag{20}$$

Thus, the measures used to evaluate the effects of learning are avg. LD, relevant coverage, query diversity, and EoD.

4. User study

The synthesized evaluation framework from the previous section will now be applied to measure and compare exploratory search performance among individuals, dyads, and triads. The present section provides details of a user study conducted for this purpose.

4.1 Subjects

Participants in this study were students recruited from a major US university through open calls that were spread through various e-mail-lists. Through an online form, the participants signed up for one of the following three project conditions:

- C1: individually;
- C2: in pairs (dyads); and
- C3: in a group of three (triads).

This provided with three different experimental conditions, which was the primary independent variable of this study. While signing up as a team, the participants were required to pair up with either one or two people with whom they had previously worked. This design decision was made in order to ensure that participants had common ground and to make the collaborative task more realistic. Such requirements have been reported in some of the earlier studies (Shah and González-Ibáñez, 2011; González-Ibáñez *et al.*, 2012; Su, 1992).

From the 68 participants (12 individuals, ten dyads, and 12 triads) that were recruited, 40 were female and 28 were male, with ages ranging between 18 and 24. Most of the participants (60 percent) reported using the Windows operating system. Moreover, all of the participants indicated having intermediate to advanced search skills.

The participants were compensated according to the project condition they signed up for. Those who signed up individually received $10, those in pairs received $12.50 per person, and those in a triad received $15 per person. In addition, they were informed that the best performing individual or team would receive $25 per person at the end of the study to encourage them to take the task more seriously. It was also ensured that recruits participated only once irrespective of their group membership.

4.2 Collaborative search system

A modified version of Coagmento (González-Ibáñez and Shah, 2011), an open source plugin for the Firefox Web browser, was used to provide a set of tools for supporting information seeking, sharing, synthesis, as well as communication for teams.

As depicted in Figure 1, the version of Coagmento used in this study consists of two major components: toolbar and sidebar. The toolbar contains three buttons: first, search, which provides access to Google search engine; second, snip for saving and sharing portions of texts of a given Web page; and third, editor, which opens a collaborative editor for writing the report required in the task (see description below).

The sidebar serves three primary functions: first, provides awareness of the remaining time for the task; second, displays snippets collected; and third, provides text-based communication channel.

4.3 Logging tools

Beyond the features for supporting collaboration among team members, Coagmento also provides powerful logging functionality capable of recording users' actions within

Figure 1.
A snapshot of the experimental system with parts of it shown in details

the Firefox Web browser, which includes every Web page visited, queries run on any Web search engine as well as within specialized sites such as Wikipedia, and snippets collected by highlighting text on a Web page and pressing "Snip" button on Coagmento toolbar.

All of this data were recorded with timestamps. In addition to recording the Web browsing activity of users, Coagmento was also used to record the messages exchanged during the collaboration process, and data from questionnaires that were introduced as part of the system.

4.4 Study setup

The study was conducted in an interaction lab at a major US university that included a participant's room and an observer's room. In the participant's room (Plate 1), three computers were organized in a triangle formation, allowing up to three participants to work with each other. In the case of one or two participants, the same one or two computers were used to keep identical spatial configurations within those conditions.

The audio and video were also recorded and streamed in real time to a screen in the next room where a researcher was stationed as shown in Plate 2. This allowed the researcher to monitor the study session without disturbing the participants. In addition, the researcher and the participants were also able to communicate when required (e.g. participants inquires and technical difficulties) via an external text-chat system.

The participants were provided with mid-range desktop computers with 20' LCD screens and running Windows 7. They were also required to use the Firefox Web browser since the Coagmento plugin at the time was available for that particular browser only. This also allowed keeping the experimental parameters consistent among all the participants. In addition to the computing resources, the participants were given a whiteboard with markers and an eraser. The temperature and humidity in the study room were maintained at steady and comfortable levels throughout all the sessions.

Plate 1.
A session with a triad in progress

Plate 2.
A researcher monitoring the participants' room from the observer's room using an audio/video monitoring system

4.5 Session workflow

Coagmento was adapted to guide users through various stages in the session workflow of this study from stage two up to stage six as shown in Table I. The researcher guided stages 1 and 7.

Stage	Description	Time (min)
1	Participants were introduced to the study and asked to sign a consent form	4
2	Participants filled in a demographic questionnaire	1
3	Participants watched a brief tutorial in order to learn the basic functionalities required during the task	3
4	Participants worked on a simple practice task to get accustomed to the system	3
5	Participants read the task description (presented above)	1
6	Participants individually filled out a set of pre-task questionnaires, which include questions about task familiarity and perceived difficulty	1
7	Each user/team worked for approximately 35 minutes on the given task that included searching and collecting relevant information, and using it to compose a report (typically last 5 minutes)	35
8	Participants filled out post-task questionnaires, which include questions about perceived difficulty and confidence in their response	2
9	Participants were interviewed briefly to get their views about the task, their experience, and feedback	5
Total		55

Table I. Summary of session stages

4.6 Task

The participants were asked to collect relevant information in an exploratory search task designed to be a realistic work-task as described in Borlund and Ingwersen (1999). The topic of "global warming" was selected for the task, which according to a few pilot runs was found to be appropriate in terms of the amount of material available on this topic and how engaging it was for the participants. The task description was presented as follows:

> A leading newspaper agency has hired your team to create a comprehensive report on the causes, effects, and consequences of the climate change taking place due to global warming. As a part of your contract, you are required to collect all the relevant information from any available online sources that you can find.
>
> To prepare this report, search and visit any Website that you want and look for specific aspects as given in the guideline below. As you find useful information, highlight and save relevant snippets. Later, you can use these snippets to compile your report, no longer than 200 lines, as instructed.
>
> Your report on this topic should address the following: Description about global warming, how it affects climate change, scientific evidence about global warming affecting climate change, causes of global warming, consequences of global warming causing climate change, measures that different countries around the globe has taken over the years to address this issue and recent advancements in addressing this issue. Also describe different view points people have about global warming (specify at least three different view points you find) and relate those to the aspects controversies held by the public on this topic.

The tasks carried out within Coagmento do not require participants to engage in a process of task initiation as described by Kuhlthau (1991) in her information search process (ISP) model, during which the information seeker recognizes the need for new information to complete an assignment, with its associated feelings of apprehension and uncertainty. The seeker is also relieved of the responsibility to select a topic, the ISP's second stage, and the presentation of work is relatively perfunctory and circumscribed. There remains, however, the requirement that participants locate relevant

information – a process requiring evaluative judgments – and that the information selected be integrated – synthesized – into a product measured against imposed criteria. This design allows for a more focussed investigation of collaborative search and synthesis processes performed in different spatial conditions. Of course, Kuhlthau's is not the only model that one could use to define or evaluate this task. Another relevant model is that of Marchionini (1995). Based on his model, the participants in the study described here covered all the sub-processes, except problem recognition, problem definition, and system selection. This should not come as a surprise since these three sub-processes were predetermined for the participants as a part of the study design.

5. Results

The analyses of the data collected from the user study were primarily divided in two categories: individual user level and team level. This was required to understand the differences among the users as well as projects in different conditions. While several kinds of data were collected during the study, the analyses presented here focusses on interpreting the log data. Also, for the purpose of these analyses, the report-writing portion that occurred at the end of the session will be ignored since the focus here is on the search episode. As mentioned in the previous section, the independent variable of this study was group size: individuals, dyads, triads, leading to the three conditions: C1, C2, and C3.

According to *Q-Q* plots, histograms, and Shapiro-Wilk test, data for each measure within each condition was found to be non-normal. Results from Brown-Forsythe showed that the assumption of homogeneity of variance was not violated. Therefore, the Kruskall-Wallis test was used to perform comparisons across conditions. *Post hoc* tests were performed with the Wilcoxon rank-sum test when the Kruskall-Wallis test result was significant at $p < 0.01$. Note that since only medians are reported here, aggregated values may not give the overall picture. Also, since the Kruskall-Wallis and Wilcoxon tests were conducted using medians, means, or standard deviations are not reported.

The results of log data mining on user-level data are presented in Table II, with the results of statistical tests summarized in Table IV. Given that a primary interest here is to evaluate exploratory search done in collaboration, user-level measures alone may not be appropriate. The results of log data analysis on team-level data are also presented in Table III, with the results of statistical tests summarized in Table IV. The results are organized according to the five measurement categories presented in the framework described in evaluation framework section.

5.1 Information exposure

As evident from the user-level results (Tables II and IV), there were no differences among the users of the three conditions for visited Web pages (coverage). Those in C1, however, had higher values for recall, but it turns out that those in C2 had a higher score for LD than C1 and those in C3 outperformed both C1 and C2 in this measure. In other words, dyads were able to find more information that was hard to find and triads were even better at this.

While at user level there were no differences for coverage and unique coverage, those in C2 at the team level were exposed to more information than C1. Those in C3 were exposed to more information than both C1 and C2 (Tables III and IV). The user-level and team-level results for information exposure indicate that the teams (C2 and C3) were able to divide up the work appropriately and individually explore information without much overlap with their teammates.

Measures	C1	C2	C3
Information exposure			
Coverage	17	11	13
Unique coverage	5	3.5	5
Avg. likelihood of discovery	−0.01	−0.01	−0.01
Recall	0.05	0.03	0.03
Information relevancy			
Relevant coverage	11	7	7
Unique relevant coverage	2	1	2
Num saved (snippets)	15	9	9
Precision	0.35	0.23	0.21
Information search			
Num distinct queries	9	5.5	6
Query diversity	19.59	18.26	16.60
Avg query Info content	2.51	2.52	2.50
SERP	13	7	9.5
Performance			
F-score	0.08	0.05	0.05
Effectiveness	0.43	0.42	0.50
Efficiency	0.05	0.07	0.08
Learning			
Topic familiarity (pre-task)	4	3	3
Perceived challenge (pre-task)	3	3	3
Perceived difficulty (post-task)	4	5	4
Report confidence (post-task)	3	3	3

Table II. Medians per condition for various measures with user as the unit of analysis

5.2 Information relevancy

Looking at the relevance of the information the participants were exposed to as shown in Table II, it seems that while the individuals (C1) had better relevant coverage, there were no differences in this measure that was unique for a given condition (unique relevant coverage). Individually, C1 users were also able to collect more information than those in other conditions.

However, looking at the team-level evaluations, as depicted in Table III, teams outperform individuals in almost all the measures. Teams (C2 and C3) were not only able to achieve more coverage that was deemed pertinent (relevant coverage), but were also able to discover information that no other units in project conditions (individuals or dyads) found. unique relevant coverage here is an example of how simply looking at the median values (Table III) does not provide the correct sense of differences between conditions. For reference, looking at the means, C1 had 2.42 (SD = 2.27), C2 had 2.7 (SD = 1.83), and C3 had 5.67 (SD = 2.46) for this measure. Using Kruskall-Wallis and Wilcoxon tests, it was observed that there were more values in C2 that were above the median compared to C1. Similarly, C3 had more values above the median compared to both C1 and C2.

5.3 Information search

No differences were found in individual user information search behaviors, as reflected by various query-based features, across the three conditions, except for SERP, which represents how many of the search engine results were visited. Once again, while C1

Measures	C1	C2	C3
Information exposure			
Coverage	17	25	38
Unique coverage	5	10	17.50
Avg. likelihood of discovery	−0.01	−0.01	−0.01
Recall	0.05	0.07	0.09
Information relevancy			
Relevant coverage	11	15	21
Unique relevant coverage	2	2	5.50
Num saved (snippets)	15	17.50	26
Precision	0.35	0.21	0.19
Information search			
Num distinct queries	9	15	16.50
Query diversity	19.59	22.54	23.46
Avg query info content	2.51	2.53	2.50
SERP	13	14	26.50
Performance			
F-score	0.08	0.10	0.12
Effectiveness	0.43	0.43	0.47
Efficiency	0.05	0.03	0.02
Learning			
Topic familiarity (pre-task)	4	3	3
Perceived challenge (pre-task)	3	3	3
Perceived difficulty (post-task)	4	5	4
Report confidence (post-task)	3	3	3

Table III.
Medians per condition for various measures with team as the unit of analysis

had a higher score for this measure, C2 outperformed C1, and C3 outperformed both C1 and C2 at the project level. In other words, individuals (C1) accessed more SERPs per person than those working in collaboration (C2 and C3), but at the project level, those in collaborative conditions were able to get to more SERPs than those working individually. This shows the effectiveness of division of labor in C2 and C3. While a higher quantity in overall number of SERPs seen could show the effectiveness of an ISP, the quality of this process is unclear. There is, however, a naive way to at least think about it. Given that each unit was allotted the same amount of time to work on the proposed task, those in C1 spent the least amount of time per SERP per person than those in C2 and C3. In other words, participants in collaborative conditions spent more time assessing the results per person (leading to possibly higher quality), while achieving higher quantity of SERPs at the project level. The evidence of this quantity-quality relationship with respect to individual user-level vs team-level activities could be confirmed by looking at relevant coverage as described in Section 5.2.

5.4 Performance

Not surprisingly, at the user level, C1 achieved higher *F*-scores than C2 and C3, given that C1 had higher recall than the other two conditions and there were no differences for precision. However, both C2 and C3 outperformed C1 with regard to efficiency. Once again, looking at project-level results, the teams were found to be performing better than the individuals (i.e. C3 > C2 > C1).

Measures	User level	Team level
Information exposure		
Coverage	=	C2 > C1 C3 > C1,C2
Unique coverage	=	C2 > C1 C3 > C1,C2
Avg. likelihood of discovery	C2 > C1 C3 > C1,C2	=
Recall	C1 > C2,C3 C2 > C3	C2 > C1 C3 > C1,C2
Information relevancy		
Relevant coverage	C1 > C2,C3 C2 > C3	C2 > C1 C3 > C1,C2
Unique relevant coverage	=	C2 > C1 C3 > C1,C2
Num saved (snippets)	C1 > C2,C3 C3 > C2	C2 > C1 C3 > C1,C2
Precision	=	C1 > C2,C3 C2 > C3
Information search		
Num distinct queries	=	C2 > C1 C3 > C1,C2
Query diversity	=	=
Avg query info content	=	=
SERP	C1 > C2,C3 C3 > C2	C2 > C1 C3 > C1,C2
Performance		
F-score	C1 > C2,C3 C3 > C2	C2 > C1 C3 > C1,C2
Effectiveness	=	=
Efficiency	C2 > C1,C3 C3 > C1	=
Learning		
Topic familiarity (pre-task)	=	
Perceived challenge (pre-task)	=	
Perceived difficulty (post-task)	=	
Report confidence (post-task)	=	

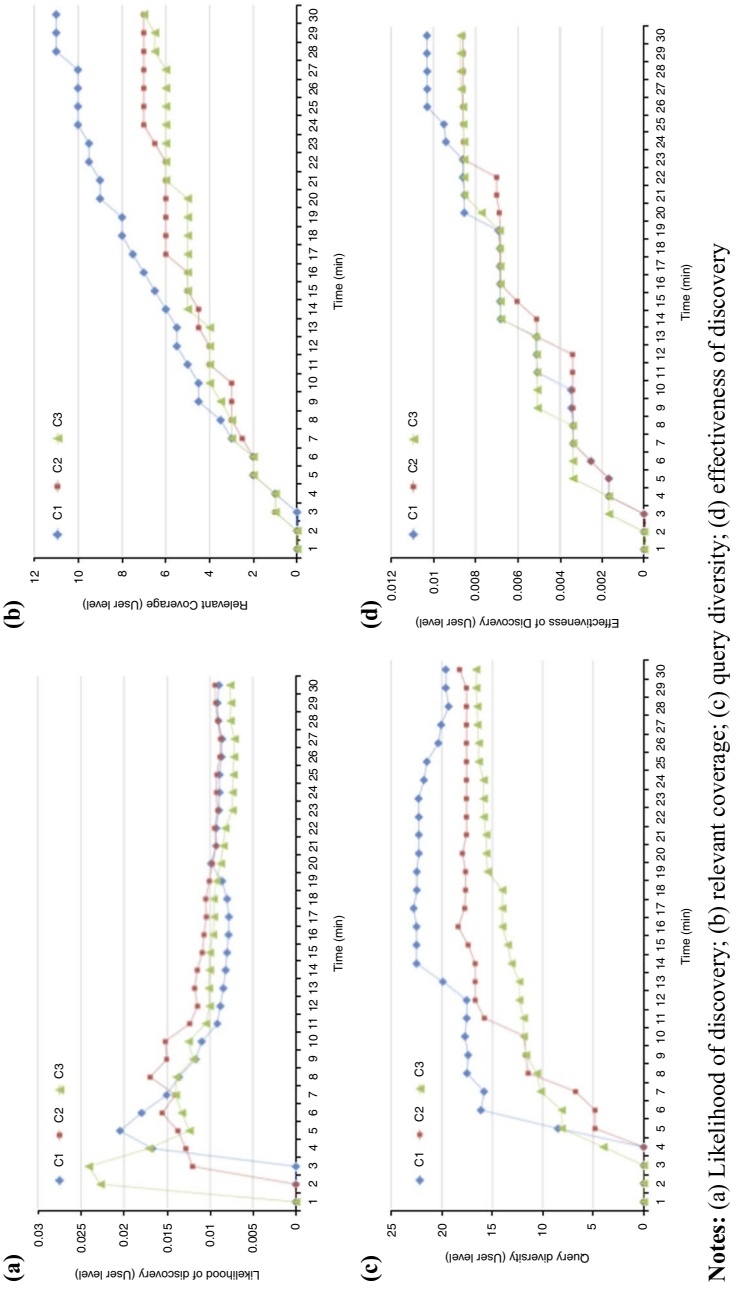
Note: Differences at $p < 0.01$

Table IV. Results of Kruskall-Wallis and Wilcoxon rank-sum test as *post hoc* with individual users and teams as the unit of analysis

5.5 Learning

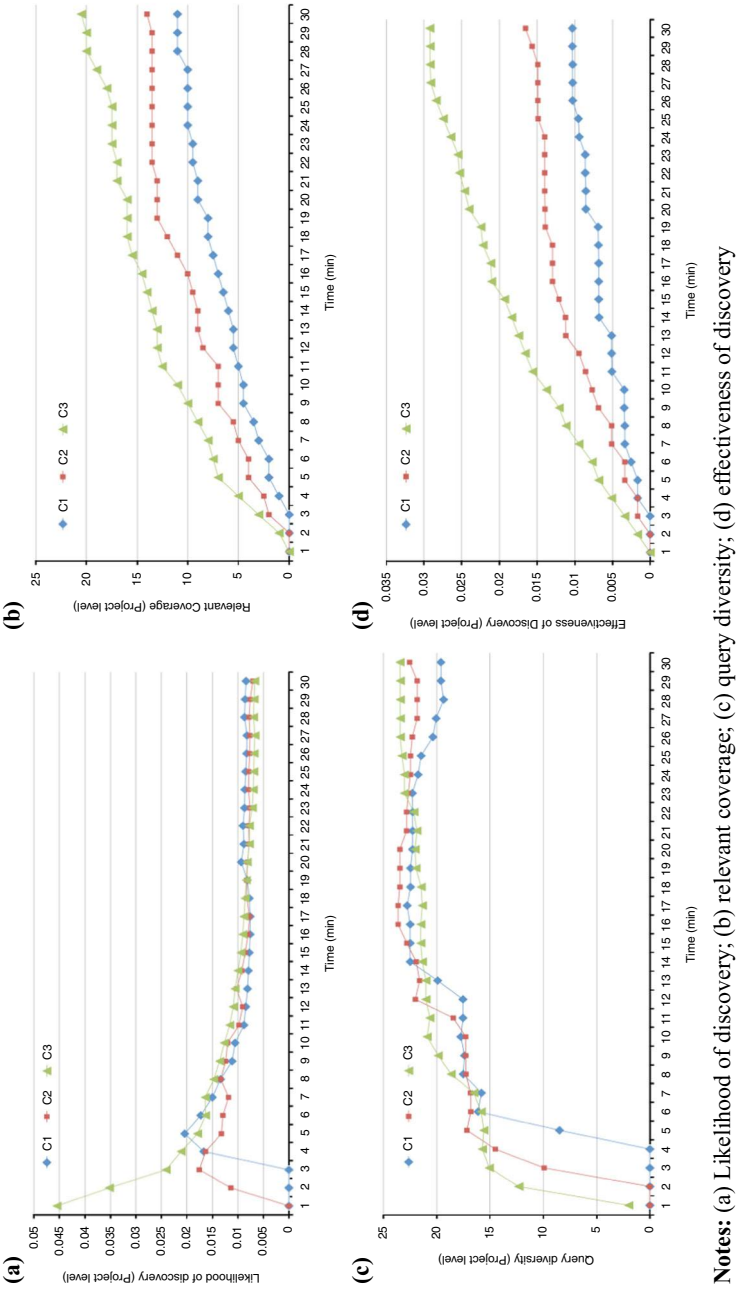
By analyzing participants' responses to pre-task and post-task questions, no differences were found with regard to topic familiarity or perceived challenge at the moment of first being exposed to the task description. In a similar way, at the end of the task, no differences were found in terms of perceived difficulty or confidence in the report created as part of the task. However, these questions do not necessarily provide a comprehensive view of learning that took place during the session. Therefore, log data were used to test four learning hypotheses using the measurements described in Section 3.5.

The results of applying these measures are depicted in Figure 2 (user level) and Figure 3 (project level). Figures 2(a) and 3(a) show that in the beginning, the



Notes: (a) Likelihood of discovery; (b) relevant coverage; (c) query diversity; (d) effectiveness of discovery

Figure 2.
Time-series for user-level measures



Notes: (a) Likelihood of discovery; (b) relevant coverage; (c) query diversity; (d) effectiveness of discovery

Figure 3.
Time-series for team-level measures

users/teams discovered information that is easy to find (higher values for LD), but with time, they started locating Web pages that were harder to discover. Figures 2(b) and 3(b) show the increase in covering relevant information with time. Similarly, Figures 2(c) and 3(c) indicate that with time the users/teams were trying more diverse queries. Finally, Figures 2(d) and 3(d) provide evidence of users'/teams' increased effectiveness in discovering information in a constantly shrinking universe of information.

Not surprisingly, analyzing the relationship between the measures calculated over time per condition by performing trend analysis using linear regression fitting with intercept (where the dependent variable is the value of the respective measure at each time point and the independent variable is the time in minutes) confirms significant linear relationships across all of these measures. Tables V and VI summarize these results, indicating percent of instances that showed significant relationships at $p < 0.05$. While this is not a perfect measure for evaluating the learning that occurred, these measures do provide evidence for the effect of learning, and therefore, reject all the null hypotheses ($H_0 1$-$H_0 4$) presented earlier.

6. Discussion

It should be clear from the background, the evaluation framework, the results, and the analyses of a user study that evaluating exploratory Web search is a challenging task, whether such searching is done individually or collaboratively. In most cases, such an evaluation must employ several measures and rely on analyses that go beyond system-focussed computations of measures such as recall and precision, as well as user-focussed calculations of usability. Having recognized these challenges, the present paper attempted to provide a framework that uniquely combines various forms of log data collected through a typical user study involving exploratory search. The work reported here also expanded this evaluation to include collaborative search scenarios that highlight additional challenges involved in comparing individual and collaborative search projects.

Table V.
Portion of instances at user level that showed significant relationships in trend analysis as a way to indicate learning

Measures	C1 (%)	C2 (%)	C3 (%)
Likelihood of discovery	66.67	50.00	66.67
Effectiveness of discovery	100.00	100.00	100.00
Query diversity	91.67	100.00	97.22
Relevant coverage	100.00	100.00	100.00

Note: $p < 0.05$

Table VI.
Portion of instances at team level that showed significant relationships in trend analysis as a way to indicate learning

Measures	C1 (%)	C2 (%)	C3 (%)
Likelihood of discovery	66.67	50.00	91.67
Effectiveness of discovery	100.00	100.00	100.00
Query diversity	91.67	80.00	100.00
Relevant coverage	100.00	100.00	100.00

Note: $p < 0.05$

This approach also reveals some of the limitations of the work reported here. First, the evaluation framework assumes certain forms of data (log data from Web searches), nature of the experiments (interactive IR study with users), and a bound on time and available information. Second, the proposed evaluation framework is synthesized using relevant literature on exploratory and collaborative exploratory search that may omit other views and approaches. Third, the study reported here demonstrates how the proposed evaluation framework could be used in a lab setting and exhibits several constraints (time-bound task, artificial motivation, fixed configuration of tools and space, to name a few) by design. A more realistic experiment, where the participants have a strong connection or motivation for doing the task, may present a different set of challenges and results. Fourth, the relatively small sample size (10-12 per units of analysis per condition) is acknowledged. This sample size, unsurprisingly, was set due to practical limits imposed by resources (time and money) available. However, given several clear patterns in the findings, a larger sample is likely to yield similar results. Finally, several factors that may affect the recommendation and/or evaluation of collaborative searching are not explicitly considered here. These include communication, cost of collaboration, cognitive load (Fidel *et al.*, 2004), and affective load. However, if the number of queries used and the number of Web pages visited indicate a level of physical effort at the user level, results show that users in dyads and triads did not exert additional effort with these information-related measures. This indicates that groups were able to successfully divide up the work in the exploratory search task and take advantage of collaboration without incurring additional costs. Participants' responses to the questionnaire based on NASA's task load index instrument (Hart and Staveland, 1988), which elicits cognitive effort experienced during a task, also revealed no significant differences among the participants in different conditions.

Despite the limitations reported above, the work presented here provides interesting insights into evaluating individual and collaborative exploratory Web search. For instance, it was found that those working in collaboration are able to locate some information that participants working alone could not discover. In fact, the triads performed even better than dyads in this regard. Figure 4, inspired by the work of Shah and González-Ibáñez (2011) and drawn to scale, depicts this insight. Various forms of coverage for each condition are indicated using overlapping rectangles within an encompassing area that represents overall coverage. As shown, there is a large portion of the available information that C3 teams discovered that C1 and C2 did not (unique coverage for C3). This has implications for general Web search. It is often found that individuals do not go beyond the first few results (usually the first SERP) when doing a Web search, thus missing out on some relevant and/or novel information that may be available lower in the rank-list. Having two or more people work together without requiring any additional changes to their search behavior easily enabled them to split the task in such a way that they end up discovering those areas of the information landscape that they might have otherwise missed working by themselves.

7. Conclusion

To investigate exploratory search performances for individuals as well as teams of two and three, a user study was conducted. Since there is no one or simple way to evaluate exploratory search, an evaluation framework was synthesized using relevant literature. Through the application of this framework on the user study data,

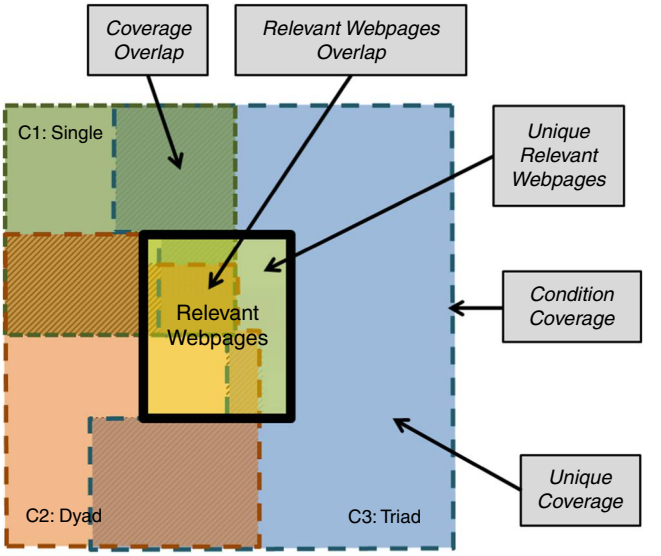


Figure 4.
Depiction of coverage by various conditions

Source: Shah and González-Ibáñez (2011) (drawn to scale)

it was observed that at the user level, users in teams and individual users did not exhibit any significant difference among the measures calculated for information exposure, information relevance, information search, and performance. In contrast, when evaluated at the team level, dyads and triads (C2 and C3) outperformed individuals (C1) in most measures in information exposure, information relevance, information search, performance, and learning. For the aspect of learning, dyads and triads were able to achieve improved results over time compared to individuals in almost all measures, showing that as a team they were able to learn more about the task and gather the required information at a faster pace.

These results support the following conclusions: evaluating exploratory search – for individuals or collaborators – should incorporate measuring various forms of activities, objects, and processes; to evaluate collaborative IR, one needs to apply these measures at both the user level and the team level; and given an appropriate setup, collaborators could outperform individuals in an exploratory search task through the synergic effect (Shah and González-Ibáñez, 2011).

While it was found that two people are better than one and three collaborators are better than two, it remains to be explored how many more collaborators one could add to a team before reaching the point of diminishing returns. It seems that C3 teams were able to perform well in terms of coverage and search effectiveness due to their ability to successfully divide up the work in three parts while exploring subspaces of the information landscape with minimal overlap. However, if the number of participants were greater than the number of aspects/facets of the topic, a team may not be able to apply this strategy, leading to saturation or even worsen their performance. This hypothesis is worth testing for future work.

Note

1. See (Shah, 2014a) for a detailed review of the current literature.

References

Allan, J. (2004), "HARD track overview in TREC 2004: high accuracy retrieval from documents", in *Proceedings of TREC 2004*, pp. 24-37.

Allen, B. (1997), "Information needs: a person-in-situation approach", *Proceedings of an International Conference on Information Seeking in Context. ISIC '96. Taylor Graham Publishing, London*, pp. 111-122, available at: http://dl.acm.org/citation.cfm?id=267190.267197 (accessed October 13, 2015).

Aneiros, M. and Estivill-Castro, V. (2005), "Usability of real-time unconstrained www-co-browsing for educational settings", *Proceedings – 2005 IEEE/WIC/ACM International Conference on Web Intelligence, WI, 2005*, pp. 105-111.

Baeza-Yates, R. and Pino, J.A. (1997), "A first step to formally evaluate collaborative work", *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work: the Integration Challenge – GROUP '97*, pp. 56-60, available at: http://dl.acm.org/citation.cfm?id=266838.266860 (accessed October 13, 2015).

Bates, M.J. (1979), "Information search tactics", *Journal of the American Society for Information Science*, Vol. 30 No. 4, pp. 205-214, available at: http://dx.doi.org/10.1002/asi.4630300406

Bawden, D. (1986), "Information systems and the stimulation of creativity", *Journal of Information Science*, Vol. 12 No. 5, pp. 203-216.

Belkin, N.J., Marchetti, P.G. and Cool, C. (1993), "BRAQUE: design of an interface to support user interaction in information retrieval", *Information Processing and Management*, Vol. 29 No. 3, pp. 325-344.

Borlund, P. and Ingwersen, P. (1999), "The application of work tasks in connection with the evaluation of interactive information retrieval systems: empirical results", *Proceedings of the 1999 International Conference on Final Mira. MIRA'99. UK: British Computer Society, Swinton*, pp. 1, available at: http://dl.acm.org/citation.cfm?id=2228065.2228066 (accessed October 13, 2015).

Broder, A. (2002), "A taxonomy of web search", *SIGIR Forum*, Vol. 36 No. 2, pp. 3-10.

Budd, J.M. (2004), "Relevance : language, semantics, philosophy", *Library Trends*, Vol. 52 No. 3, pp. 447-462.

Denning, P.J. (2007), "Mastering the mess", *Communications of the ACM*, Vol. 50 No. 4, pp. 21-25.

Denning, P.J. and Yaholkovsky, P. (2008), "Getting to 'we' ", *Communications of ACM*, Vol. 51 No. 4, pp. 19-24.

Fidel, R., Pejtersen, A.M., Cleal, B. and Bruce, H. (2004), "A multidimensional approach to the study of human-information interaction: a case study of collaborative information retrieval", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 11, pp. 939-953.

Foster, A. and Ford, N. (2003), "Serendipity and information seeking: an empirical study", *Journal of Documentation*, Vol. 59 No. 3, pp. 321-340.

Foster, J. (2006), "Collaborative information seeking and retrieval", *Annual Review of Information Science and Technology*, Vol. 40 No. 1, pp. 329-356.

Fox, S., Karnawat, K., Mydland, M., Dumais, S. and White, T. (2005), "Evaluating implicit measures to improve web search", *ACM Transactions on Information Systems*, Vol. 23 No. 2, pp. 147-168.

González-Ibáñez, R. and Shah, C. (2011), "Coagmento: a system for supporting collaborative information seeking", *Proceedings of the American Society for Information Science and Technology*, Vol. 48 No. 1, pp. 1-4, available at: http://doi.wiley.com/10.1002/meet.2011.14504801336

González-Ibáñez, R., Haseki, M. and Shah, C. (2012), "Time and space in collaborative information seeking: the clash of effectiveness and uniqueness", *Proceedings of the American Society for Information Science and Technology*, Vol. 49 No. 1, pp. 1-10, available at: http://dx.doi.org/10.1002/meet.14504901080

González-Ibáñez, R., Haseki, M. and Shah, C. (2013), "Let's search together, but not too close! An analysis of communication and performance in collaborative information seeking", *Information Processing & Management*, Vol. 49 No. 5, pp. 1165-1179.

González-Ibáñez, R., Shah, C. and White, R.W. (2012), "Pseudo-collaboration as a method to perform selective algorithmic mediation in collaborative IR systems", *Proceedings of the Association of Information Science & Technology (ASIST) Annual Meeting, Baltimore, MD*.

Hansen, P. and Järvelin, K. (2005), "Collaborative information retrieval in an information-intensive domain", *Information Processing and Management*, Vol. 41 No. 5, pp. 1101-1119.

Hart, S.G. and Staveland, L.E. (1988), "Development of NASA-TLX (task load index): results of empirical and theoretical research", in Hancock, P.A. and Meshkati, N. (Eds), *Advances in Psychology*, Vol. 52, North-Holland, pp. 139-183.

Hyldegård, J. (2006), "Collaborative information behaviour-exploring Kuhlthau's information search process model in a group-based educational setting", *Information Processing and Management*, Vol. 42 No. 1, pp. 276-298.

Jansen, B.J., Booth, D.L. and Spink, A. (2008), "Determining the informational, navigational, and transactional intent of web queries", *Information Processing and Management*, Vol. 44 No. 3, pp. 1251-1266.

Kuhlthau, C.C. (1991), "Inside the search process: information seeking from the user's perspective", *Journal of the American Society for Information Science*, Vol. 42 No. 5, pp. 361-371.

Kules, B. and Capra, R. (2009), "Designing exploratory search tasks for user studies of information seeking support systems", *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries. JCDL '09. ACM, New York, NY*, pp. 419-420.

Marchionini, G. (1995), *Information Seeking in Electronic Environments*, Cambridge University Press, New York, NY.

Marchionini, G. (2006), "Exploratory search: from finding to understanding", *Communications of the ACM*, Vol. 49 No. 4, pp. 41-46.

Morris, M. (2007), "Collaborating alone and together: investigating persistent and multi-user web search activities", *Proceedings of international ACM SIGIR Conference, Amsterdam, Netherlands, SIGIR'07, July 23-27*.

Morris, M.R. (2008), "A survey of collaborative web search practices", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1657-1660.

Morris, M.R., Paepcke, A. and Winograd, T. (2006), "TeamSearch: comparing techniques for co-present collaborative search of digital media", *Proceedings of the First IEEE International Workshop on Horizontal Interactive Human-Computer Systems, TABLETOP'06*, pp. 97-104.

Pickens, J. and Golovchinsky, G. (2007), "Collaborative exploratory search", In Proc 2007 HCIR Workshop, pp. 21-22, available at: www.fxpal.com/?p=CES.

Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P. and Back, M. (2008), "Algorithmic mediation for collaborative exploratory search", *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval – SIGIR '08*, pp. 315, available at: http://portal.acm.org/citation.cfm?doid=1390334.1390389 (accessed October 13, 2015).

Reddy, M.C. and Jansen, B.J. (2008), "A model for understanding collaborative information behavior in context: a study of two healthcare teams", *Information Processing and Management*, Vol. 44 No. 1, pp. 256-273.

Rose, D.E. and Levinson, D. (2004), "Understanding user goals in web search", *Proceedings of the 13th international conference on World Wide Web (WWW '04). ACM, New York, NY*, pp. 13-19, available at: http://doi.acm.org/10.1145/988672.988675 (accessed October 13, 2015).

Saracevic, T. (1995), "Evaluation of evaluation in information retrieval", *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '95. ACM, New York, NY*, pp. 138-146, available at: http://doi.acm.org/10.1145/215206.215351 (accessed October 13, 2015).

Saracevic, T. (2007), "Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 1915-1933.

Shah, C. (2014a), "Collaborative information seeking", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 2, pp. 215-236.

Shah, C. (2014b), "Evaluating collaborative information seeking – synthesis, suggestions, and structure", *Journal of Information Science*, Vol. 40 No. 4, pp. 460-475.

Shah, C. and González-Ibáñez, R. (2011), "Evaluating the synergic effect of collaboration in information seeking", *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information – SIGIR '11*, pp. 913-922.

Shah, C. and González-Ibáñez, R. (2012), "Spatial context in collaborative information seeking", *Journal of Information Science*, Vol. 38 No. 4, pp. 333--349.

Shah, C. and Marchionini, M. (2010), "Awareness in collaborative information seeking", *Journal of American Society of Information Science and Technology*, Vol. 61 No. 10, pp. 1970-1986.

Shah, C., Pickens, J. and Golovchinsky, G. (2010), "Role-based results redistribution for collaborative information retrieval", *Information Processing & Management*, Vol. 46 No. 6, pp. 773-781.

Smyth, B., Balfe, E., Boydell, O., Bradley, K., Briggs, P., Coyle, M. and Freyne, J. (2005), "A live-user evaluation of collaborative web search", *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1419-1424.

Sonnenwald, D.H. (1999), "Evolving perspectives of human information behaviour: contexts, situations, social networks and information horizons", in Wilson, T.D. and Allen, D.K. (Eds), *Exploring the Contexts of Information Behaviour: Proceedings of the 2nd. International Conference on Research in Information Needs, Seeking and Use in Different Contexts, Taylor Graham, London*, pp. 176-190.

Sonnenwald, D.H. and Pierce, L.G. (2000), "Information behavior in dynamic group work contexts: interwoven situational awareness, dense social networks and contested collaboration in command and control", *Information Processing and Management*, Vol. 36 No. 3, pp. 461-479.

Soulier, L., Shah, C. and Tamine, L. (2014), "User-driven system-mediated collaborative information retrieval", *Proceedings of ACM Special Interest Group on Information Retrieval (SIGIR), Gold Coast*, pp. 485-494, available at: http://dl.acm.org/citation.cfm?id=2609598 (accessed October 13, 2015).

Spink, A. (2002), "A user-centered approach to evaluating human interaction with web search engines: an exploratory study", *Information Processing & Management*, Vol. 38 No. 3, pp. 401-426.

Su, L.T. (1992), "Evaluation measures for interactive information retrieval", *Information Processing & Management*, Vol. 28 No. 4, pp. 503-516.

Talja, S. (2002), "Information sharing in academic communities: types and levels of collaboration in information seeking and use", *New Review of Information Behavior Research*, Vol. 3, pp. 143-159.

White, R.W. and Huang, J. (2010), "Assessing the scenic route : measuring the value of search trails in web logs" search", pp. 587-594, available at: http://portal.acm.org/citation.cfm?id=1835548 (accessed October 13, 2015).

White, R.W. and Roth, R.A. (2009), "Exploratory search: beyond the query-response paradigm", *Synthesis Lectures on Information Concepts, Retrieval, and Services*, Vol. 1 No. 1, pp. 1-98.

White, R.W., Muresan, G. and Marchionini, G. (2006a), "Report on ACM SIGIR 2006 workshop on evaluating exploratory search systems", *SIGIR Forum*, Vol. 40 No. 2, pp. 52-60.

White, R.W., Kules, B., Drucker, S.M. and Schraefel, M.C. (2006b), "Supporting exploratory search", *Communications of the ACM*, Vol. 49 No. 4, pp. 37-40.

Wildemuth, B.M. and Freund, L. (2012), "Assigning search tasks designed to elicit exploratory search behaviors", *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval – HCIR '12, (C)*, pp. 1-10.

Wilson, M.L. and schraefel, M.C. (2008), "Evaluating collaborative search interfaces with information seeking theory", available at: http://eprints.soton.ac.uk/265669/ (accessed October 13, 2015).

Wilson, T.D. (1981), "On user studies and information needs", *Journal of Documentation*, Vol. 37 No. 1, pp. 3-15.

Corresponding author
Dr Chirag Shah can be contacted at: chirags@rutgers.edu