Evaluating the Synergic Effect of Collaboration in Information Seeking

Chirag Shah
School of Communication & Information (SC&I)
Rutgers, The State University of New Jersey
4 Huntington St, New Brunswick, NJ 08901, USA
chirags@rutgers.edu

Roberto González-Ibáñez
School of Communication & Information (SC&I)
Rutgers, The State University of New Jersey
4 Huntington St, New Brunswick, NJ 08901, USA
rgonzal@eden.rutgers.edu

ABSTRACT
It is typically expected that when people work together, they can often accomplish goals that are difficult or even impossible for individuals. We consider this notion of the group achieving more than the sum of all individuals' achievements to be the synergic effect in collaboration. Similar expectation exists for people working in collaboration for information seeking tasks. We, however, lack a methodology and appropriate evaluation metrics for studying and measuring the synergic effect. In this paper we demonstrate how to evaluate this effect and discuss what it means to various collaborative information seeking (CIS) situations. We present a user study with four different conditions: single user, pair of users at the same computer, pair of users at different computers and co-located, and pair of users remotely located. Each of these individuals or pairs was given the same task of information seeking and usage for the same amount of time. We then combined the outputs of single independent users to form artificial pairs, and compared against the real pairs. Not surprisingly, participants using different computers (co-located or remotely located) were able to cover more information sources than those using a single computer (single user or a pair). But more interestingly, we found that real pairs with their own computers (co-located or remotely located) were able to cover more unique and useful information than that of the artificially created pairs. This indicates that those working in collaboration achieved something greater and better than what could be achieved by adding independent users, thus, demonstrating the synergic effect. Remotely located real teams were also able to formulate a wider range of queries than those pairs that were co-located or artificially created. This shows that the collaborators working remotely were able to achieve synergy while still being able to think and work independently. Through the experiments and measurements presented here, we have also contributed a unique methodology and an evaluation metric for CIS.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'11, July 24–28, 2011, Beijing, China.
Copyright 2011 ACM 978-1-4503-0757-4/11/07...$10.00.

Categories and Subject Descriptors
H.3: INFORMATION STORAGE AND RETRIEVAL **H.3.3: Information Search and Retrieval**: *Search process*; **H.5.3** [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Collaborative computing, Computer-supported cooperative work*

General Terms
Experimentation; Human Factors; Measurement

Keywords
Collaborative information seeking; Synergic effect; Evaluation.

1. INTRODUCTION
It is often argued that people should work together to solve hard problems (e.g., [1,2]). Information seeking is, mistakenly, not seen as a tough problem that could be effectively solved through collaboration [9,10]. Instead, IR researchers tend to spend their efforts on improving the systems that could boost an individual's search effectiveness. While this approach has achieved remarkable feat as evident by systems, theories, and their effects reported in the literature, including that of this forum, it is time to revisit the notion of collaboration and social/interactive elements of searching. In this paper we show how and why people working in collaboration for an information-seeking task could do far better than those working individually. More specifically, we look at the notion of synergy, where the whole becomes more than the sum of all, and demonstrate that an appropriate setup of collaborative information seeking (CIS) could lead to covering information that could otherwise not be discovered by individuals. We also show how remotely located collaborators achieve synergy by exercising independence and diversity.

In the next section, we review some of the relevant literature on CIS and collaboration in general focusing on two specific aspects: synergy and evaluation. Our specific research questions and hypotheses are also included in this section. To address these questions and specifically the hypotheses, we conducted a laboratory study with 10 single users, and 30 pairs of users in three different setups. The details of this study are given in Section 3. Following that, we describe what and how we evaluated in Section 4. Using these evaluation measures, we present the results and relevant discussion in Section 5. The paper is concluded in Section 6 with implications and directions for further research.

2. BACKGROUND

A brief overview of some of the related research is presented here to situate our work in the context of IR in general and CIS in particular. This review will also inform our research questions and hypotheses listed later in this section.

2.1 Related work

Synergy, from the Greek *synergos* (which means working together) is when two or more things, individuals, disciplines, etc. interact producing a result that is greater than the sum of the individual products of the involved parts [1]; this is commonly explained through the expression *1+1>2*. Synergy has been widely studied in a variety of contexts, which includes, chemistry, medicine, organizations, and education [1]. In the particular context of collaborative information seeking (CIS), little is known about the synergic effect and to what extent, if any, it affects users' behaviors, the processes, and the results produced by individuals searching information in collaboration, in comparison to those working individually. Furthermore, collaboration has been conceptually defined in different ways and to date there is no clear difference between it and related concepts such as cooperation [21]. Authors such as Fidel et al. [4]; Golovchinsky et al. [9], and Shah [21] have further defined and explored the conceptual implications of collaboration in order to have a more suitable definition for CIS.

Unlike the study of information seeking performed by single users, little is known about how users seek information in collaboration with others. During the past decade, some have explored different aspects of CIS in both naturalistic and experimental settings. The focus, however, has been on describing users' behaviors as well as the information seeking processes of teams. For example, Hyldegard [12, 13] studied the applicability of Kuhlthau's information search process [15] in the context of groups. Similarly, Shah & González-Ibáñez [22] attempted to map the stages in Kuhlthau's model to collaborative information seeking. Both studies revealed that even though some stages of this model may apply to CIS, they do not cover the social dimension of CIS.

More importantly, it has been recognized that collaborative search is a phenomenon widely present in social context where collaboration - in a more general sense - is required. For example, Morris [17] conducted a survey in a large company that revealed that workers engage in a variety of activities in which they collaborate searching for information. Through this survey, the author also identified obstacles that impact the collaborative information seeking process, different tasks that motivate collaborative search, and common methods that are employed to share search results during the process.

Specific studies have been conducted with the aim of comparing individual and collaborative search. For instance, Joho, Hannah, & Joemon [14] compared the search process of single users with collaborative-concurrent search. Through this study, the authors found that those working collaboratively were able to reduce overlapping in terms of the webpages covered during a recall-oriented search task; however, as the authors pointed out, this redundancy reduction did not improve the retrieval effectiveness. In a related study, Pickens et al. [19] and Shah et al. [23] found that collaborative search, with algorithmic mediation to enhance the collaboration process among participants, end up with better results than those obtained by merging of single users' results. Similarly, Foley [7] demonstrated that in order to enhance the performance in synchronous collaborative information retrieval (SCIR), it is necessary to have an appropriated division of labor and also a mediated support for sharing knowledge. On a more theoretical side, González-Ibáñez [10] proposed a conceptual map to study the behavior of users in the information seeking process of both individuals and collaborative teams. As the author suggested, such a map and the use of mixed methods would help to understand different dimensions involved in CIS, including the synergic effect of collaboration.

Though early studies have suggested the importance of providing support for collaboration in situations involving information search practices [25], and many have devised tools and technologies to facilitate CIS [16, 18, 26]; it has been recognized by several authors (e.g., [5, 20]) that there is a significant lack of theories and tools to investigate and support CIS. Moreover, to our knowledge, beyond the coverage of these and other works, none have studied specifically the synergic effect in collaborative information seeking, which is one of the core values and advantages of collaboration.

2.2 Research questions and hypotheses

From reviewing the literature, it is clear that (1) while collaboration is often considered a useful approach to solve a complex problem, it has been largely ignored when it comes to information seeking; (2) system-focused CIS works have primarily looked at creating algorithms and tools that provide improved effectiveness in collaboration, but ignore the actual user interaction and social component; and (3) interface-focused CIS works have predominantly studied various usability issues with the interface/system in a CIS situation. These explorations have not addressed some of the basic questions about working in collaboration, including the following that we are interested in investigating here.

1. How does collaborating with a shared workstation differ from that of working with individual computers?
2. How does the location of collaborators affect their interactions and effectiveness in a CIS task?
3. In a recall-oriented exploratory search task, can we simply combine the outputs of two independent users to achieve the same/similar results as a pair to produce the same results as users actually working together?
4. How can we measure the synergic effect in collaboration and what does it inform us?
5. To what extent, if any, does the cognitive load of collaborative teams working in a search task differs from the cognitive load of a single user working in the same task?

To narrow our investigation for better articulation of our goals and experiments, we propose the following hypotheses. Note that they are postulated for a recall-oriented exploratory search task.

1. Two collaborators working at the same computer will be able to accomplish as much as that of a single user.
2. Remotely located collaborators will be able to work more independently compared to those working co-located.
3. Real collaborative teams will accomplish more compared to artificially created teams by combining the outputs of two independent users.

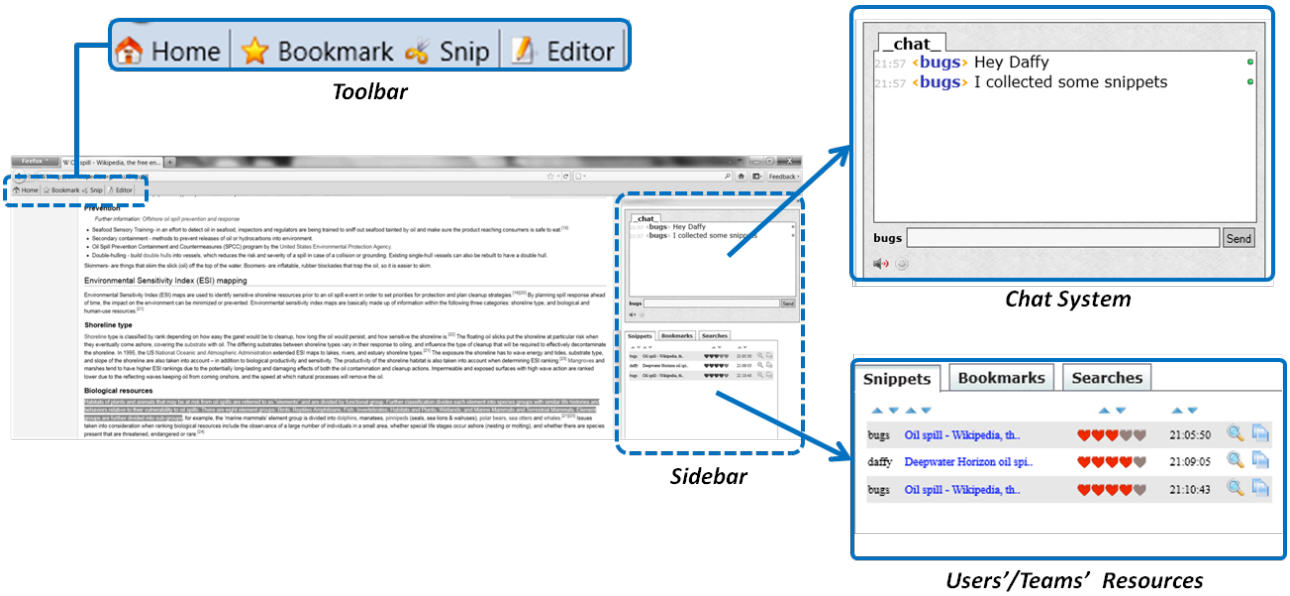

Figure 1: A snapshot of the experimental system with parts of it shown in details.

The following section describes a user study that we designed to test these hypotheses and address the research questions listed before.

3. METHOD

We conducted a laboratory study involving a total of 70 participants – 10 participants as single users, and 60 participants as collaborative teams. This section describes the study procedure, the subjects, the system, the task, and the experimental conditions.

3.1 Subjects

Participants in this study were asked to sign up in pairs with someone with whom they had previous experience collaborating. In addition, they were informed of their compensation for participating in the study, which consisted of $10 per person and the possibility to obtain additional prizes if they were among the three best performing teams ($50, $25, and $15 per person additionally) at the end of the study. Overall, 60 participants in 30 pairs, all of them students from Rutgers University, were recruited and randomly assigned to three experimental scenarios (Table 1). In addition, 10 more participants were recruited to work individually in the same search task performed by teams.

3.2 System

We developed a plugin for the Firefox web browser, called Coagmento,[1] which provided appropriate tools and support for the participants working in various conditions. A screenshot of this plugin within Firefox is shown in Figure 1. As shown, the plugin included a toolbar and a sidebar. The toolbar had the following buttons: (1) Home – for taking the participant to appropriate questionnaires, (2) Bookmark – for bookmarking a webpage, (3) Snip – for collecting a snippet using highlighted text from a webpage, and (4) Editor – for accessing a shared editor for writing the report.

The sidebar had two major components: a chat-box and a resources panel. The chat-box allowed the collaborators in a given team to communicate with each other. The researcher conducting the study also used it to provide instructions to the participants. The resources panel included tabs for bookmarks, saved snippets, and executed queries.

3.3 Session workflow

Each experimental session lasted less than an hour and was structured in six parts as depicted in Figure 2 and described below.

1. Participants were introduced to the study and asked to sign a consent form.
2. Participants watched a brief tutorial in order to learn the basic functionalities required during the task.
3. Participants individually filled out a set of pre-task questionnaires. In the case of two participants working at the same computer (condition later explained), the participants were separated for this phase.
4. Participants read the task description (given later).
5. Each participant/team worked for approximately 25 minutes on the given task that included searching for

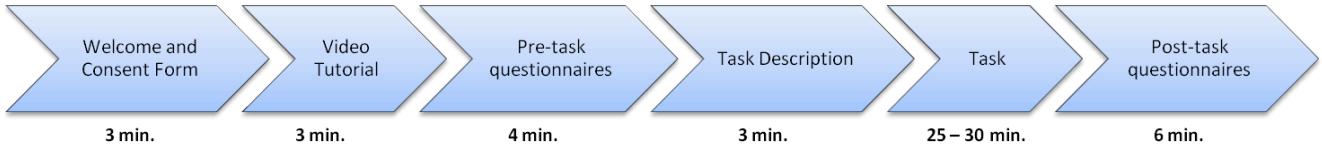

Figure 2: Study session workflow.

[1] Publicly available from http://www.coagmento.org.

relevant information, and using it to compose a report.

6. Participants filled out post-task questionnaires.

The researcher conducting the study communicated with the participants through the chat-box at different times during the study instructing them to start/stop the task or fill in a questionnaire.

3.4 Conditions

To study the difference between individual information seeking and CIS, as well as to understand how various CIS settings can affect a collaborative team's effectiveness in accomplishing an information-seeking task, we conducted experiments with four different conditions: single participants, two participants at the same computer, two participants in the same room but different computers, and two participants in different rooms with individual computers.

In order to have a baseline to study the synergic effect of collaboration, we artificially created pairs of users from C1 (single users). We generated all possible combinations of pairs in groups of 5, reaching a total of 49 groups and creating 245 artificial teams in total. This was done in order to cover all possible pairs of users while avoiding a given user appearing in more than one team within the same group of teams.

These five conditions are summarized in Table 1. Setups for four of these conditions are also depicted in Figure 3. Note that in the real experiment, those in C5 condition were located in different rooms separated by walls, and not just a partition. They could not see or talk to each other directly, and the only communication channel they had was the text-box provided with the system.

Table 1: Experimental conditions.

Cond.	Description
C1	Single participants
C2	Artificial team
C3	Co-located using the same computer
C4	Co-located using different computers
C5	Remotely located

3.5 Task

We chose "gulf oil spill" as the topic for this experimentation since it was quite popular and relevant at the time the study was being conducted. Our preliminary investigations, including a few pilot runs, indicated that there was a huge amount of material on this topic, and that the participants would find it interesting and challenging enough as an exploratory search task. Each participant was given the following task description.

"A leading newspaper has hired your team to create a comprehensive report on the causes, effects, and consequences of the recent gulf oil spill. 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. Make sure you also rate a snippet to help you in ranking them based on their quality and usefulness. 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 issues: description of how the oil spill took place, reactions by BP as well as various government and other agencies, impact on economy and life (people and animals) in the gulf, attempts to fix the leaking well and to clean the waters, long-term implications and lessons learned."

The participants saw this description on the screen (phase 4 in the study), and were also given a printed copy to refer to during their session.

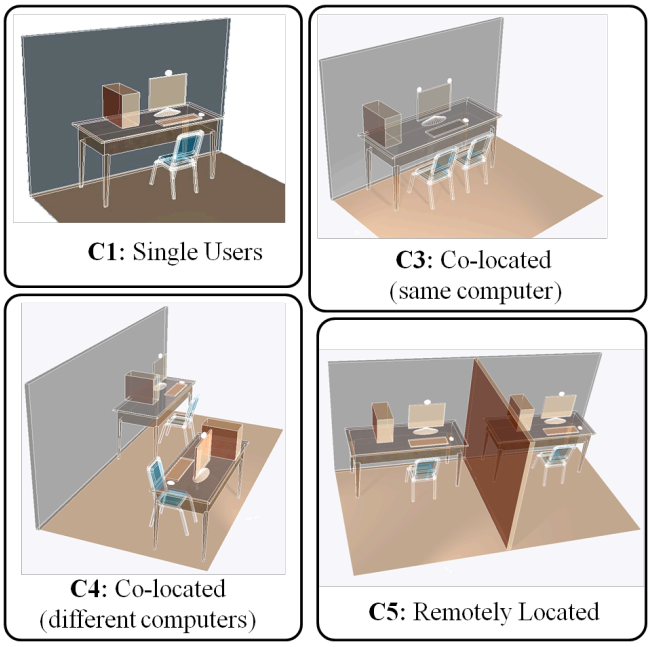

Figure 3: Experimental setups for four different conditions.

4. EVALUATION

In order to evaluate the effectiveness of the participants in various conditions, we employed a number of traditional and non-traditional evaluation measures, which are presented below. Here we also describe other useful constructs and definitions that will later be used while reporting and discussing the results.

4.1 Universe of webpages

In order to compute quantities such as coverage, we needed a universal set of webpages. Given that the search domain for our experiments was the open web, we needed a more confined set that we could use to compare with. We decided to take the union of all the webpages visited by all of our participants (total 70). In other words, the universe of webpages was defined by combining the visited webpages of each participant/team in every condition.

$$U = \bigcup_{t} Coverage(t) \qquad \ldots (1)$$

Here, *Coverage(t)* is the coverage (webpages visited) by participant/team *t*.

4.2 Relevant webpages

This corresponds to the webpages that participants either bookmarked using the toolbar or from where one or more snippets were collected. Once again, we took the union of all such webpages by each participant/team to form a universe of relevant webpages.

$$U_r = \bigcup_t RelevantCoverage(t) \qquad ... \ (2)$$

Here, *RelevantCoverage(t)* is the set of webpages that participant/team *t* visited and found as relevant.

4.3 Precision, recall, and F-measure

Two of the most common evaluation measures in IR are precision and recall, which for our purpose here, are defined as the following.

$$Precision(t) = \frac{RelevantCoverage(t)}{Coverage(t)} \qquad ... \ (3)$$

$$Recall(t) = \frac{RelevantCoverage(t)}{U_r} \qquad ... \ (4)$$

To combine precision and recall into one measure of effectiveness, we use the traditional formulation of F-measure as defined below.

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \qquad ... \ (5)$$

4.4 Coverage

We defined coverage of a given team/participant as the total number of distinct webpages visited within the universe of webpages.

$$Coverage(t) = \{wp_i : wp_i \ was \ visited \ by \ t \wedge wp_i \in U\} \qquad ... \ (6)$$

We also considered a particular region of the coverage of teams/participants that was unique within the universe. We called such region unique coverage, which consists of all webpages within the coverage of a given team/participant *t* that were visited only by *t*.

$$UniqueCoverage(t) = Coverage(t) \setminus \bigcup_{t_i \in (T \setminus \{t\})} Coverage(t_i) \ ...(7)$$

4.5 Relevant coverage

We defined relevant coverage as the region of coverage of a given team/participant that intersects with the universe of relevant webpages.

$$\text{RelevantCoverage}(t) = \text{Coverage}(t) \cap \text{U}_\text{r} \qquad ... \ (8)$$

In a similar way, we called unique relevant coverage to the set of webpages within the unique coverage of a given team/participant that intersects with the universe of all relevant webpages.

$$\text{UniqueRelevantCoverage}(t) = \text{U}\textit{nique}\text{Coverage}(t) \cap \text{U}_\text{r} \qquad ... \ (9)$$

4.6 Useful webpages

In addition to relevance, we also studied the usefulness of webpages that teams/participants visited during the task. We used an implicit measure based on the dwell time on a webpage as described in [27], which is supported by previous findings [8]. As reported in these prior works, we considered a webpage to be useful if a team/user spent at least 30 seconds on it. Using the log data, we computed dwell time on a given webpage by a user, and if it was greater than or equal to 30 seconds, marked it as useful for that user. Note that we only considered content pages, discounting any search engine homepage or search engine results pages (SERPs).

4.7 Likelihood of discovery

To evaluate effectiveness of a team/participant in discovering hard to find information, we devised a new measured, called *likelihood of discovery*. We assume that webpages with a high likelihood are easier to find and are common among the majority of the users. On the other hand, those webpages with a low likelihood are difficult to reach and probably beyond the first results page of search engines. A participant/team finding these webpages are being more effective in discovering information that is not just relevant, but also diverse.

In order to operationalize this idea, we used a formulation similar to that of inverse document frequency (IDF). Using the frequency of each webpage in our log data, we computed its likelihood to be visited; in addition, we multiplied each webpage's likelihood by -1 in order to denote the IDF. As a result, each webpage was assigned with a normalized value between -1 and 0. In this sense, those webpages with a value close to 0 are rare (and even unique) to be reached by teams/participants, while those close to -1 are more likely to be visited.

4.8 Query diversity

In addition to the sources that teams/participants visited during the tasks, we also studied how they approached the task in terms of the queries they issued to find information. We studied how similar or different were the queries formulated by participants in real teams and artificial teams.

In order to evaluate query diversity, we used Lavenshtein distance to compute the distance between pairs of queries for each real team and also for all combinations of users in artificial teams. Based on the results of this computation, for a given pair of queries; the closer the distance to 0 the higher the similarity between them. On the other hand, the higher the distance between queries, more different (therefore diverse) were the queries formulated within a team.

4.9 Cognitive Load

To study if collaboration has some negative implications for users in terms of cognitive load experienced; we asked our participants to respond a questionnaire after finishing their task. This questionnaire was a simplified version of NASA's Task Load indeX (TLX) [6].[2] This instrument had the following questions.

1. *How mentally demanding was this task?*
2. *How physically demanding was this task?*
3. *How hurried or rushed was the pace of the task?*
4. *How hard did you have to work to accomplish your level of performance?*
5. *How insecure, discouraged, irritated, stressed, and annoyed were you?*

All these questions were responded using a 5-point Likert scale, where higher values in the responses indicate a more negative perception of the user with respect to of the areas considered in the above questions.

[2] Taken from http://www.cc.gatech.edu/classes/AY2005/cs7470 fall/papers/manual.pdf

5. RESULTS AND DISCUSSION

In this section we present our results and their related discussions. To facilitate this, we first provide a summary of all the universal sets used in Table 2. These sets were used to compute other constructs, such as precision, recall, and unique relevant coverage. Note that the numbers in this table represent the combined output of all the 70 participants.

Table 2: Summary of various universes used in our analysis.

	Total
Universe of all webpages (U)	562
Universe unique webpages	377
Universe all relevant webpages (U_r)	228
Universe unique relevant webpages	159

The analyses reported in this section were done using one-way ANOVA. We tested for homogeneity of variance and performed appropriate post-hoc tests to measure the difference between the conditions.

5.1 Precision, recall, and F-measure

To begin our analysis, we first looked at simple precision and recall for each condition as defined in the previous section. A summary of these measures is given in Table 3, with variance analysis in Table 4.

We found no difference between any of the conditions for precision. This is not surprising considering how it was computed and that it was relatively easy to find relevant results from the web, giving almost everyone a very high value for precision. This high precision was also due to the fact that the relevant set was constructed using the union of the relevance judgments provided by each participant.

Table 3: Relevance measures – means and standard deviations for each condition.

	Cond.	1	2	3	4	5
Precision	**Mean (s.d.)**	0.704 (0.293)	0.646 (0.233)	0.682 (0.141)	0.532 (0.132)	0.552 (0.163)
Recall	**Mean (s.d.)**	0.053 (0.032)	0.104 (0.046)	0.058 (0.032)	0.113 (0.038)	0.129 (0.040)
F-measure	**Mean (s.d.)**	0.092 (0.049)	0.165 (0.054)	0.106 (0.054)	0.182 (0.054)	0.202 (0.051)

We did, however, find differences among the conditions for recall. Not surprisingly, the single users (C1) had lower recall compared to every other condition except C3, where two collaborators used the same computer. Similarly, C3 had lower recall than C2, C4, or C5. In other words, teams with individual computers for their collaborators were able to achieve higher recall than those with a single or a shared computer. This was expected since the assigned task was recall-oriented and exploratory in nature, and given the limited amount of time, those with more resources achieved more results. If the task was non-dividable (e.g., brainstorming), we may not have found these differences.

Using F-measure, once again, we found that in real collaboration, those with individual computers (C4 and C5) outperformed those with shared computers (C3).

Table 4: Means differences between conditions (row-column). Bold values correspond to statistical significant difference at p≤.05 and italic bold values at p≤.01.

	Cond.	2	3	4	5
Precision	**1**	0.058	0.021	0.172	0.152
	2		-0.036	0.113	0.094
	3			0.150	0.131
	4				-0.019
Recall	**1**	***-0.051***	-0.005	**-0.059**	***-0.075***
	2		**0.046**	-0.008	-0.025
	3			**-0.054**	***-0.070***
	4				-0.016
F-measure	**1**	***-0.073***	-0.013	***-0.090***	***-0.109***
	2		***0.059***	-0.016	-0.036
	3			**-0.076**	***-0.096***
	4				-0.019

5.2 Coverage

While precision and recall correspond to relevance in traditional IR sense, they are not very appropriate in the present CIS setting given that the participants were searching the web, where a huge amount of information on the given topic existed, and that each participant/team was given the same amount of limited time, in which they could easily find a good amount of relevant information. Given this, a more important and interesting aspect to investigate here is coverage – a measure of the amount of information explored. Tables 5 and 6 report various kinds of coverage quantities, including the analysis of variance between different conditions. Just as recall, we found those with two people and two computers were able to cover more information than those with only one person and/or one workstation.

To extend our investigation for coverage, we looked at unique coverage for each individual/team, which is defined as the set of unique webpages that one covered and others did not. We found that C4 and C5 came out on the top with this measure, indicating their effectiveness in covering information that others could not. In fact, C5 outperformed not only C1 and C3, but also C2. In other words, when two real collaborators worked in remote CIS,

they were able to cover more unique information than artificially created pairs of collaborators. Given that C2 and C4 had no difference, we can say that there is a value in remote collaboration when the task has clear independent components, and at the same time, having interactions that one finds in a real collaboration helps over completely working independently as C2 participants did.

We also looked at how much of the unique coverage was actually relevant, and found that C5 did better than C1 and C3. While we found no difference between C2 and C4 or C5, we can clearly see that those in C5 were able to get to the information that is not discovered by single users (C1), collaborators at the same computer (C3), or artificially created collaboration (C2). At the same time, the amount of unique relevant information that they discovered was found to be significantly more than what was found by C1 or C3. In other words, remotely located teams were able to leverage real interactions (as opposed to no interactions in artificial teams), and at the same time carry on independent exploration to achieve the synergic effect through collaboration.

Table 5: Coverage of information – means and standard deviations for each condition.

	Cond.	1	2	3	4	5
Coverage	**Mean (s.d.)**	8.50 (5.148)	16.57 (7.247)	9.30 (5.056)	17.90 (6.118)	20.50 (6.329)
Relevant Coverage	**Mean (s.d.)**	4.80 (2.098)	9.21 (2.323)	6.40 (3.658)	9.60 (4.377)	11.10 (4.458)
Unique Coverage	**Mean (s.d.)**	3.60 (3.062)	7.20 (4.707)	3.80 (3.225)	11.00 (3.712)	12.20 (5.007)
Unique Relevant Coverage	**Mean (s.d.)**	1.30 (1.418)	2.60 (1.766)	1.80 (1.687)	3.40 (1.506)	3.90 (1.729)

Figure 4 provides a depiction of various forms of coverage for C2, C3, C4, and C5. The figure is drawn to scale, with the area of a coverage region, proportional to the value of that kind of coverage for a given condition. Visually also, we could see that C5 has more coverage (total and unique), as well as more unique relevant coverage.

5.3 Usefulness

Moving beyond information exploration and relevancy, we looked at the usefulness of viewed information. As defined in the previous section, this referred to visiting webpages that one spends considerable amount of time (30 seconds or more).

We found (Tables 7 and 8) that C1 and C3 visited significantly fewer useful webpages than those by C4 and C5 participants. More importantly, real teams with individual computers (C4 and C5) outperformed artificial teams (C2) when it came to visiting useful webpages, including the webpages that were unique to a given team. In other words, with real collaboration, the teams were able to avoid overlapping their exploration and reach out to more sources of information. This is important for exploratory and recall-oriented search task like the one used for our study.

Table 6: Means differences between conditions (row-column). Bold values correspond to statistical significant difference at p≤.05 and italic bold values at p≤.01.

	Cond.	2	3	4	5
Coverage	**1**	***-8.067***	-0.8	**-9.4**	***-12.0***
	2		**7.267**	-1.333	-3.933
	3			**-8.6**	***-11.2***
	4				-2.6
Relevant Coverage	**1**	***-4.412***	-1.6	**-4.8**	***-6.3***
	2		2.812	-0.388	-1.888
	3			-3.2	-4.7
	4				-1.5
Unique Coverage	**1**	-3.6	-0.2	***-7.4***	***-8.6***
	2		3.4	-3.8	***-5.0***
	3			***-7.2***	***-8.4***
	4				-1.2
Unique Relevant Coverage	**1**	-1.3	-0.5	**-2.1**	***-2.6***
	2		0.8	-0.8	-1.3
	3			-1.6	**-2.1**
	4				-0.5

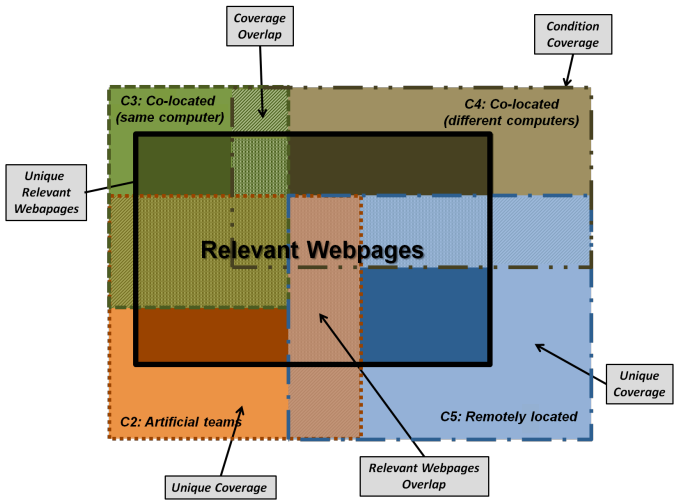

Figure 4: Depiction of coverage by various conditions (drawn to scale).

Table 7: Usefulness of explored information – means and standard deviations for each condition.

	Cond.	1	2	3	4	5
Useful Webpages	**Mean (s.d.)**	4.50 (2.593)	9.00 (3.726)	6.10 (3.315)	13.50 (4.007)	13.60 (2.914)
Unique Useful Webpages	**Mean (s.d.)**	1.90 (1.792)	3.80 (2.799)	2.10 (2.132)	7.50 (2.718)	7.30 (2.983)
Likelihood of discovery	**Mean (s.d.)**	-0.011 (0.005)	-0.008 (0.003)	-0.009 (0.006)	-0.006 (0.002)	-0.005 (0.001)

Table 8: Means differences between conditions (row-column). Bold values correspond to statistical significant difference at p≤.05 and italic bold values at p≤.01.

	Cond.	2	3	4	5
Useful Webpages	1	***-4.5***	-1.6	***-9.0***	***-9.1***
	2		2.9	***-4.5***	***-4.6***
	3			***-7.4***	***-7.5***
	4				-0.1
Unique Useful Webpages	1	-1.9	-0.2	***-5.6***	***-5.4***
	2		1.7	***-3.7***	***-3.5***
	3			***-5.4***	***-5.2***
	4				0.2
Likelihood of discovery	1	-0.0039	-0.0021	**-0.0056**	**-0.0064**
	2		-0.0017	-0.0017	**-0.0024**
	3			-0.0035	-0.0042
	4				-0.0007

In addition, we found that C4 and C5 participants visited many more difficult to reach (less likely visited) webpages than those in C1. This shows that collaboration, in this case, allowed a unit (a pair or users) get to the information that was otherwise ignored by those who worked alone. In fact, even when we combined two independent participants' explorations (C2), those who collaborated while remotely located (C5) could discover more information. Once again, C5 participants leveraged on the interactions through real collaboration while exercising their independence and exploring individual information trails.

5.4 Query diversity

To understand searching behavior of various individuals and teams, we looked at their querying effectiveness. Given that the assigned task was exploratory in nature, that there was plethora of useful information on the web, and that the time was limited, it was important that the participants construct a variety of queries and cover as large a ground as they could. To measure this, we computed query diversity for each participant/team as defined in the previous section (Table 9). A general overview of query distances for each condition is given in Figure 5. As we can see, those in C4 had more queries with higher distances among them. In other words, those in C5 tried more diverse queries.

Table 9: Query diversity – means and standard deviations for each condition.

	Cond.	2	4	5
Query Diversity	**Mean (s.d.)**	20.41 (9.205)	19.69 (8.863)	23.13 (8.465)

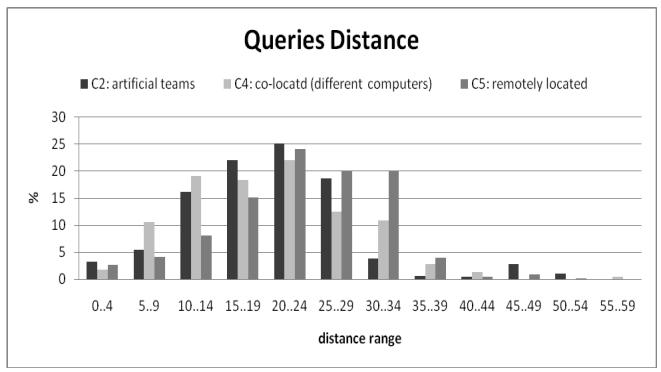

Figure 5: Query distance for C2, C4, and C5 using Lavenshtein algorithm.

Using ANOVA, we found (Table 10) that those in real collaborations with individual computers and remotely located (C5) had higher diversity in their queries than those co-located using different computers (C4). C5 participants also exhibited a larger variety in their queries than those by artificially created teams (C2). Combining these two facts indicates that (1) participants remotely located were able to successfully divide the task up and explore unique information through different queries, and (2) these participants had a chance to work more independently than those in the same space.

We believe that query similarity for teams in C4 is due to space sharing, which may influence the way in which users formulated their queries. Even though most teams split up the task, the physical closeness of users could enable them to hear what their peers think aloud; or even have brief conversations (facilitated by

face-to-face interaction). This may have influenced implicitly common queries between those participants.

Table 10: Means differences between conditions (row-column). Bold values correspond to statistical significant difference at p≤.05 and italic bold values at p≤.01.

	Cond.	2	3	4	5
Query Diversity	2			0.720	***-2.724***
	4				***-3.444***

On the other hand, we think that users working remotely located and using only text chat for communicating with each other (C5) were more able to generate different queries because they were less directly influenced by their peers during the task.

5.5 Cognitive Load

Fidel et al. [5] showed that collaboration induces additional cognitive load, what they referred to as the *collaborative load*. Often, this is the price to pay for gaining the advantages of collaborating. We used NASA's TLX instrument to measure user perceived cognitive load during the task (see the previous section). For analysis, we combined the responses obtained from six different questions and created an index, since these responses were found to be statistically reliable for this instrument. A summary of what we found for the four real conditions (C2 was artificially created) is given in Table 11.

Table 11: Cognitive load – means and standard deviations for each condition.

	Cond.	1	3	4	5
Cognitive Load	Mean (s.d.)	15.00 (2.449)	16.00 (2.714)	17.60 (3.050)	15.70 (3.028)

Table 12: Means differences between conditions (row-column). Bold values correspond to statistical significant difference at p≤.05 and italic bold values at p≤.01.

	Cond.	3	4	5
Cognitive Load	1	-1.00	-2.60	-0.70
	3		-1.60	0.30
	4			1.90

Using this index, we performed the ANOVA and found no difference between the four conditions (Table 12). This indicates that the participants who worked in collaboration (C3, C4, and C5) experienced no more cognitive load than what was reported by those in C1. In other words, the collaborators in our experiment (at least C4 and C5) were able to gain the advantages of creating synergy without additional mental load.

6. CONCLUSION

Collaboration is often a useful approach for solving a complex problem, but it has its costs and overheads [6]. One typically gets involved in collaboration if it has good benefit-to-cost ration, or if the given problem is too difficult to be solved without collaboration [21]. It is, however, difficult to measure if and how a collaborative endeavor would pay off. Traditional objective or quantitative approaches used in IR are insufficient, and subjective or qualitative approaches may be expensive or difficult to employ. In this paper we proposed and demonstrated a unique framework for evaluating various aspects of collaborative information seeking (CIS), especially the synergic effect.

Using a user study with 70 participants working on an information-seeking task in different setups, we showed how various evaluation measures could be defined and operationalized. Here we summarize our findings and link them back to the research questions (RQ) and hypotheses (HT) listed in Section 2.2.

We first argued that the traditional measures of evaluating relevance are not appropriate for such situations, and proposed either modified or new kinds of evaluations. These included coverage, usefulness, likelihood of discovery, and query diversity. We believe this itself is an important contribution to the community, helping the researchers evaluate and design CIS systems and interfaces (RQ4). The results of the experiments reported here indicated that in a recall-oriented exploratory search task, two collaborators working at the same computer achieve similar results to the individual users (RQ1, HT1). It also became clear that those in remote collaboration were able to work more independently than those that were co-located (HT2). Independence is considered an important characteristic of a successful collaboration [24, 21]. Our results also provided a strong support for synergic effect in remotely located collaborators. In particular, we showed that two people working in collaboration (C4 and C5) is not the same as having the outcomes of two completely independent individuals combined (C2); they do better in terms of discovering more and diverse information for an information-seeking task. Not only that, but the cognitive load in a real collaborative situation was found to be no more than what was perceived by those working individually (RQ5). Thus, the synergic effect of *the whole being greater than the sum of all* was demonstrated and evaluated (RQ3, HT3).

We also want to point out a few of the limitations of our experiments. The study reported here was conducted with synchronous CIS task. The findings may be impacted if the collaborators were working asynchronously. Due to the nature of the laboratory study, we also gave limited amount of time to the participants. It has been shown that collaborations lasting longer and done over multiple sessions may produce significantly different results and user experiences [21, 22]. Having more than two participants per team could also lead to different group dynamics, influencing the results. Despite these limitations, we

believe the methodology and the evaluation measures proposed and demonstrated here could help us further investigations of CIS with different setups, including asynchronous, non-time bound, multi-session, and non-dividable tasks, as well as collaborations that involve more than two participants. These contributions could be helpful for not just CIS, but IR in general.

7. REFERENCES

[1] Buckminster Fuller, R. (1975). SYNERGETICS Explorations in the Geometry of Thinking. http://www.rwgrayprojects.com/synergetics/synergetics.html.

[2] Corning, P. A. (1995). Synergy and self-organization in the evolution of complex systems. *Systems Research 12*(2):89-121.

[3] Denning, P. J. (2007). Mastering the mess. *Communications of the ACM*, *50*(4), 21-25.

[4] Denning, P. J., & Yaholkovsky, P. (2008). Getting to "We". *Communications of the ACM*, *51*(4), 19-24.

[5] Fidel, R., Bruce, H., Pejtersen, A. M., Dumais, S. T., Grudin, J., & Poltrock, S. (2000). Collaborative Information Retrieval (CIR). *The New Review of Information Behaviour Research*, 235-247.

[6] Fidel, R., Mark Pejtersen, A., Cleal, B., & 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*, *55*(11), 939-953. doi: 10.1002/asi.20041.

[7] Foley, C. (2008). Division of Labour and Sharing of Knowledge for Synchronous Collaborative Information Retrieval. PhD Thesis. School of Computing, Dublin City University.

[8] Fox, S., Karnawat, K., Mydland, M., Dumais, S., & White, T. (2005). Evaluating implicit measures to improve web search. ACM Transactions on Information Systems (TOIS). 23(2): 147-168.

[9] Golovchinsky, G., Pickens, J., & Back, M. (2008). A taxonomy of collaboration in online information seeking. *Proceedings of JCDL 2008 Workshop on Collaborative Exploratory Search*. Pittsburgh, PA.

[10] González-Ibáñez, R. (2010). A proposal for studying users' behaviors in collaborative information seeking through a convergence map. *GROUP 2010 Workshop on Collaborative Information Behavior*. Sanibel Island, Florida.

[11] Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 239-250). North Holland Press.

[12] Hyldegard, J. (2006). Collaborative information behaviour - exploring Kuhlthau's Information Search Process model in a group-based educational setting. *Information Processing and Management*, *42*, 276-298.

[13] Hyldegard, J. (2009). Beyond the search process - exploring group members' information behavior in context. *Information Processing and Management*, *45*, 142-158.

[14] Joho, H., Hannah, D., & Joemon, J. M. (2008). Comparing collaborative and independent search in a recall-oriented task. *Proceedings of Information Interaction in Context*. London, UK.

[15] Kuhlthau, C. C. (1991). Inside the Search Process: Information Seeking from the User's Perspective. *Journal of the American Society for Information Science and Technology*, *42*(5), 361-371.

[16] Morris, M. R., & Horvitz, E. (2007). SearchTogether: An Interface for Collaborative Web Search. *ACM Symposium on User Interface Software and Technology (UIST)* (pp. 3-12). Newport, RI.

[17] Morris, M. R. (2008). A survey of collaborative web search practices. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*. 2008:1657.

[18] Morris, M. R., Teevan, J., & Bush, S. (2008). Enhancing Collaborative Web Search with Personalization: Groupization, Smart Splitting, and Group Hit-Highlighting. *Proceedings of Computer Supported Cooperative Work (CSCW)*. San Diego, CA.

[19] Pickens, J., Golovchinsky, G., Shah, C., Qvarfordt, P., & Back, M. (2008). Algorithmic Mediation for Collaborative Exploratory Search. *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*. Singapore.

[20] Shah, C. (2009). Lessons and Challenges for Collaborative Information Seeking (CIS) Systems Developers. *GROUP 2009 Workshop on Collaborative Information Behavior*. Sanibel Island, Florida.

[21] Shah, C. (2010). Working in Collaboration - What, Why, and How? *Proceedings of Collaborative Information Retrieval workshop at CSCW 2010*. Savannah, GA.

[22] Shah, C., & Gonzalez-Ibanez, R. (2010). Exploring Information Seeking Processes in Collaborative Search Tasks. *Annual Meeting of the American Society for Information Science*. Pittsburgh, PA.

[23] Shah, C., Pickens, J., & Golovchinsky, G. (2010). Role-based results redistribution for collaborative information retrieval. *Information Processing & Management*, *46*(6), 773-781. doi: 10.1016/j.ipm.2009.10.002.

[24] Surowiecki, J. (2004). Wisdom of Crowds : Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations. Doubleday Publishing.

[25] Twidale, M. B. T., Nichols, D. M. N., & Paice, C. D. (1997). Browsing is a Collaborative Process. *Information Processing and Management*, *33*(6), 761-783.

[26] Twidale, M. B., & Nichols, D. M. (1996). Collaborative browsing and visualisation of the search process. *Proceedings of Aslib* (Vol. 48, pp. 177-182).

[27] White, R. W., & Huang, J. (2010). Assessing the scenic route: Measuring the value of search trails in web logs. *Proceedings of the Annual ACM Conference on Research and Development in Information Retrieval (SIGIR)*. Geneva, Switzerland.