Collaborative Human Computation as a Means of Information Management

Manas Tungare[1,2]**, Ben Hanrahan**[2]**, Ricardo Quintana-Castillo**[2]**,**
Michael Stewart[2]**, Manuel A. Pérez-Quiñones**[2]

[1]Google, Inc., [2]Dept. of Computer Science, Virginia Tech.
manas@tungare.name, {bhanraha, rqc, tgm, perez}@vt.edu

ABSTRACT

Information seeking in personal information collections such as email is often a solitary activity performed by the owner–user alone. However, information objects such as email that are the products of collaboration are inherently social objects. In this paper, we describe a technique, using email as an example, that exploits the actions of one's close social network to assist in one's own information seeking tasks. We note that tagging of email messages is an example of human computation, and then describe a system that enables the tags applied by one user to be shared with other recipients of the same email, thereby amortizing the cost of tagging and email management across all stakeholders. We discuss how such shared tagging contributes to common ground among the participants of a collaborative group, and may be performed with minimal global cognitive load by the sender of the message. We provide scenarios of collaborative information seeking tasks that include sub-tasks such as collaborative information management and synchronous re-finding of previously-encountered information. We wish to examine if such system support for semi-automated tagging reduces email overload for all users, and its impact on collaborative information seeking practices.

Author Keywords

Personal Information Management, Collaborative Information Seeking, Human Computation

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: Miscellaneous

INTRODUCTION

Research in Collaborative Information Seeking focuses on groups of individuals attempting to locate information. The group may consist of individuals who know each other (e.g. friends or members of a family, either collocated or remote, working synchronously or asynchronously) leading to explicit collaboration in the search process [12]. Or they may be complete strangers to one another (e.g. anonymous reviewers who contribute to recommender systems). The information over which they conduct their seeking activity may be information they already have seen, in which case the task is referred to as *refinding* [3]. In the case where collaborators are mutual strangers, the users who provide additional metadata are fully aware of the implications of their actions for other users. Reviews are not merely written to express oneself or to rant about an item or a place, but stem from the human urge to share one's knowledge with others expecting them to benefit from it.

Here, we identify a third type of collaboration, one in which the primary actor(s) do not necessarily act with the intention of social good (unlike an explicit collaborator or a review author), but act solely out of their own selfish motives of organizing their own personal information. We then examine how such individual actions form part of a distributed cognition system, where information seeking tasks are collaborative by virtue of a shared information space.

There is another notable difference in our proposed approach from other collaborative information seeking approaches described in the literature: usually, all collaborators involved have access to the same global corpus of information (likely, the World Wide Web, or other specialized corpora, such as video [14, 9]). In the technique we describe, it is not necessary for all participants to have access to the entire dataset: it is sufficient for them to have access to a *shared subset* of this information. It is only required that they have a vested interest in organizing their *personal subset* of this corpus.

Our approach combines themes from human computation and personal information management to establish a model of collaborative information seeking that has close parallels to the concept of *commensalism*. Commensalism in biological terms is defined as a relationship between two kinds of organisms in which one obtains food or other benefits from the other, but neither damages nor benefits it [15]. Similarly, in our model, we exploit the actions of one user to benefit others, without causing harm (in terms of extra work) to the first user.

Human Computation

The field of human computation [26] aims to harness the computational power locked in human brains to perform tasks that are inefficient or impossible for computers today to perform. E.g. humans are currently much more efficient in recognizing objects in images than are computers: this property has been used to design games that involve people tagging images with keywords describing them, such that a direct by-product of the game is a set of tagged images [27]. The ability of the human mind to recognize mildly-deformed text from scanned material is used in applications such as re-CAPTCHA [28], a tool that can distinguish human operators from automated robots, and at the same time digitizes text from scanned books. Techniques inspired by Human Computation are also used in Web search [17] and other general-purpose tasks via Amazon's Mechanical Turk service[1].

Personal Information Management

Personal Information is defined [16] as information that is controlled by or owned by us, about us, directed towards us, sent (posted, provided) by us, (already) experienced by us, or relevant (useful) to us. Studies in Personal Information Management seek to understand various types of personal information (e.g. files, calendars, contacts, etc.), approaches and user traits (pilers versus filers, browsing versus searching, etc.) and cross-project information management. Research in locating information that has already been encountered by a user is termed as *refinding*, and has been studied widely by Teevan [25], Capra [3], and others. Tagging as a means of information management has also been studied [21], including vocabulary issues and user incentives.

COLLABORATIVE HUMAN COMPUTATION

In this paper, we describe a technique that uses the implicit human computation performed by a user's social network to assist in his/her information management. We use email as an example information collection that illustrates our approach; however, the same general idea can be extended to several domains that involve collaboration among a close-knit set of individuals, and an information need that can be fulfilled by a shared corpus.

Email is Social, not just Personal Information

Information objects such as email are, by their very nature, artifacts of collaboration and communication. Thus, almost by definition, email messages are social objects. Via our prototype, we attempt to bring the advantages of collaborative, social information seeking practices to partly-personal-partly-social information objects such as email.

Several email studies have shown that despite excellent search tools, many users prefer to file their email into folders [30]. Recent email services provide users the option to tag emails instead of moving them to specific folders, thus providing the advantage of being able to apply two or more tags to the same message.

The premise of our design is this: collaborators who often communicate about similar topics, e.g. work colleagues, or

a close group of friends, apply similar sets of tags to their email messages. Due to the implicit one-user-one-inbox model of email, the tags applied by one user to their email are not visible to another user who may have received the exact same email from the same sender. If these tags could be shared among users who are recipients of the same email, the burden of tagging the corpus can be amortized over the entire group of collaborators. This can be done securely without compromising privacy, as described later.

The design is best communicated by discussing a scenario: (see Figure 1 for a graphical illustration.)

Scenario

Alice, Bob and Charlie are members of a research group and often collaborate in their work. Alice is a frequent filer [30] by nature and prefers to organize her email by tagging it for easy retrieval. Bob files his email irregularly, perhaps every semester. Charlie prefers not to organize his email, and instead relies on keyword searches to locate specific messages. Bob sends a message to Alice and Charlie about a Call-For-Papers and suggests that they should write about their current project, Social Email.
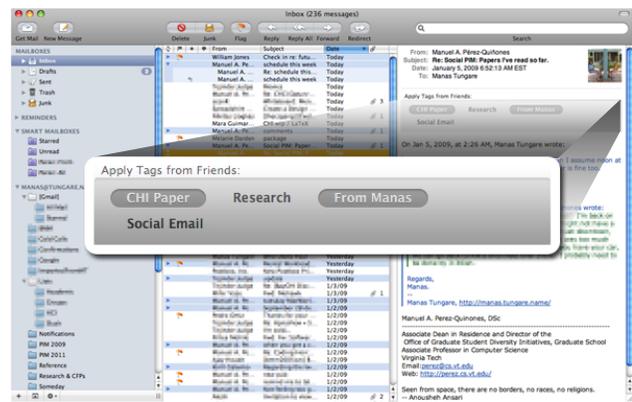

Figure 2. In-situ Tagging Interface, in Mail.app

Alice, upon receiving the email, applies the tag 'Social Email' to it, and replies saying that she is interested. If her privacy settings allow it, her tags will be sent to Bob and Charlie alongside her email (see Figure 1). When Bob and Charlie receive her reply, they see a new UI in their email client in addition to the usual display (see Figure 2). It provides them the option of tagging their email the same way Alice did. Bob agrees to let his mail client tag it as 'Social Email'. In addition, he tags it as 'Papers in Progress'. Alice and Charlie then see their TagShare UI updated to reflect the new tag applied by Bob.

Conversations consisting of multiple email messages are especially suitable for automatic tagging. Tags applied to one email of a conversation can be automatically reused for the rest of the emails of the same conversation. Any further email in that thread/conversation will have that tag already applied. This information is stored as custom email headers, thus annotating existing data as proposed by [19].
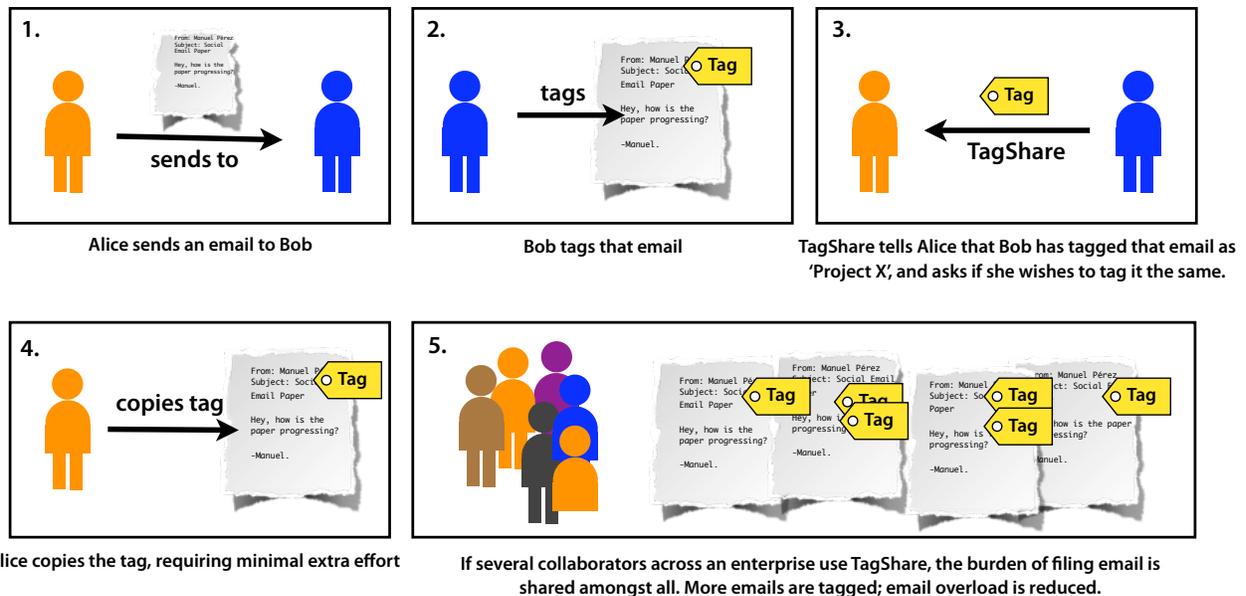
---

[1] https://www.mturk.com/

| 1. Alice sends an email to Bob | 2. Bob tags that email | 3. TagShare tells Alice that Bob has tagged that email as 'Project X', and asks if she wishes to tag it the same. |
| 4. Alice copies the tag, requiring minimal extra effort | 5. If several collaborators across an enterprise use TagShare, the burden of filing email is shared amongst all. More emails are tagged; email overload is reduced. | |

Figure 1. TagShare Usage Scenario

This is inspired by several social services that allow their users to gain from the wisdom of crowds. E.g. When bookmarking a new link on Del.icio.us, a social bookmarking service, it suggests tags applied to that same link by other users. These are often very similar to the tags a user would have applied without such prompting. The advantage gained is that users do not need an extra step of typing the tag, and thus encourages active tagging.

THEORETICAL FOUNDATIONS & PRIOR WORK
Collaborative Management of Information
The idea of collaborative management of organizational information has been pursued in several domains. Erickson [7] proposes Group Information Management as a field of inquiry that examines the semi-public sharing of personal information in social circles. He cites the example of calendaring, which has been fairly well-accepted by users, especially corporate users. Users often view and propose events to another user's calendar, thus sharing the burden of scheduling a meeting among multiple participants. SearchTogether [22] is a search environment that allows users to conduct real-time collaborative searches. Bradshaw et al. [2] propose a knowledge artifact that is the basis of information sharing and annotation among research collaborators. In a similar manner, we envision the collaborative tagging of email to be an activity that distributes the task of tagging among several willing participants who also individually stand to gain selfish benefits.

Tang et al. [24] describe their work in detecting similarities in files across the (publicly-available) home directories of corporate users. They also discuss the potential uses of the serendipitous discovery of similar files in co-workers' home directories, that may reveal shared interests. Similarly, we expect that similarities in the filing structures of corporate users (and indeed, any two collaborators) can help reveal common interests (in addition to assisting in personal email management.)

Email Management
Email users have been noted to follow different strategies. Despite the slight difference in the terminology used by various researchers, (filers and pilers [20], prioritizers and archivers [18], no-filers and filers [30], cleaners and keepers [13]), they are generally subdivided into two major camps: those who tend to keep their information organized regularly, and those who tend not to. Whittaker and Sidner identify a third tendency, and refers to such a user as a *spring cleaner* [30]. Gwizdka further notes [13] that although no-filers and spring cleaners had problems keeping up with task management via email, the advantage that filers gained was at the cost of having to spend time each day regularly in keeping their inboxes trimmed and filed.

Fisher et al. [8] describe the lessons learnt from their Social Network and Relationship Finder (SNARF). They use email history to infer an individual's social graph, and then use that to provide spatial cues for email triage. Several email providers (e.g. Gmail from Google[2]) make use of aggregate reports from their users to identify and target spam. They rely on their users clicking the "Report Spam" button to identify spam, and this knowledge is aggregated and used to mark the same email (or copies of it) as spam for other recipients as well. Garg et al. [11] describe a spam-filtering technique based on the sharing of email filters among collaborating users. Our prototype extends this to sharing not just whether a message is considered spam, but also how it was tagged/filed by one's close social network.

---

[2]http://mail.google.com/

It is important to note that the social approaches to email discussed so far use the edges from a social graph (the relationships) to help manage email. We propose enlisting the assistance of the nodes themselves (one's social contacts) to simplify email management for both users. Information from the edges of the social graph is minimally used to infer a collaborative relationship between the sender and recipient.

Tags as Common Ground

When users engage in joint actions [4], they engage in a collaborative process that creates common ground [5]. When technology is used for this collaboration, the users share either common ground through the technological artifacts [29]. The technology can help communicaiton by making the common ground available a source of referents for all parties in the collaborative activity.

In personal information management, most activity is performed in isolation. By sharing email tags amongst collaborators, our technique establishes a channel of communication that will tap the common ground that exists among participants. Since we explicitly allow the sharing of others' organizational structures, this common ground can be used to perform collaborative information seeking more efficiently.

Shared knowledge of sections of group members' information collections would facilitate the following scenario: (using the same actors as before)

Alice is on a phone conference with Bob, and references a conversation they had over email several weeks ago. Since Bob has not yet tagged his email this semester, he faces difficulty in locating the specific email. Alice remembers and notes that she had applied the tag 'Social Email' to that particular conversation, and suggests that Bob look for an email with such a tag. Bob is able to locate the email and their telephonic conversation continues, with both parties able to view the message.

The Vocabulary Problem

Furnas et al. [10] describe the vocabulary problem faced by designers in identifying a common vocabulary for their design artifacts to be shared by many of their target users. Their findings show that such a common vocabulary is difficult to find, with shared terms accounting for less than 20%. Their findings can best be summed up with this quote from their work: *"The fundamental observation is that people use a surprisingly great variety of words to refer to the same thing. In fact, the data show that no single access word, however well chosen, can be expected to cover more than a small proportion of users' attempts."*

When co-workers share work, one of the things they negotiate is the terminology they use. This terminology becomes part of their common ground [5] and it is used to make communication more effective and effortless [6]. For example, within the group of co-authors of this paper, this paper is known as the "CSCW 2010 Workshop" paper. Within the common ground of this group, that is a unique identifier that

we all agree and understand. It is often used in the subject lines of emails (e.g. "Subject: CSCW 2010 workshop submission nears"), in the body of emails (e.g. "Who is going to present the CSCW 2010 Workshop paper?"), and in phone conversations. So, although settling on a common vocabulary for all users of an arbitrary system is a problem, it is likely to be a lesser problem among smaller groups of co-workers [4]. In our study, we plan to examine whether this is indeed the case.

Also, in our proposed solution, tags are shared as they are created. This approach of sharing-early-sharing-often also prevents fragmentation of the tagspace since future taggers are aware of the tags already in use by other collaborators and are expected to be more likely to reuse those tags than to create new tags intentionally or unintentionally.

Organizing with Globally Minimal Cognitive Load

Gwizdka notes [13] that filing is a cognitively hard activity, and ideally performed as soon as an email is received. We conjecture that the sender of an email is even better equipped to tag an outgoing email with a relevant tag with the least cognitive load, because he/she is already sufficiently engaged with the material to make an informed decision. Due to the sender providing the suggested tags to the recipients of the email, the recipients will not be required to fully digest the information before making an informed tagging decision. In addition, tags applied by the sender can be consumed with lower cognitive load than the entire email itself, and can be used by the recipient to determine the relevance of an email more quickly than having to digest the entire message.

Of course, it is not a requirement that senders tag their outgoing email, but merely an optimization that is made possible by the infrastructure developed for this feature.

The Scale of Collaboration

Our initial exploration in this domain will focus on small groups within an organization for the same reasons that spam filters work well at that level, since participants already share common ground through other channels, and there is a common data store that can be used as a way to share information among them. This limits our initial approach to collaboration to within *occupational groups* [23]. It takes advantage of the common ground that exists within groups and allows users to share tags based upon their shared knowledge. Although satisfying the filing and organizing needs of individual PIM users is the primary use case of our approach, collaborative tagging and seeking practices are also well-supported.

In addition, we envision this architecture spawning new email tools that support collaborative email search tasks more directly. Within their email clients, collaborators could have functionality that enables them to connect to and share a search session, similar to the SearchTogether tool for web search [22]. Following our previous examples, Bob would have a "search with Alice" button. Upon pressing this button, they are both signed into an active search session. Alice

can help Bob find a common email and "highlight it" causing the email to pop open on Bob's screen.

Aliased Tags
We believe that the context of our proposed approach is not exactly the same as that discussed in [10]. Our proposed solution actually makes use of one of the solutions presented by Furnas et al. as Unlimited Aliasing. They state that the solution to the vocabulary problem should *'begin with user's words, and find system interpretations'*. This approach is very easily taken in our architecture: once Alice tags an email received from Bob as "Workshop", the system can propose that tag back to Bob. Bob can accept the shared tag but rename it on his end to "Paper". Note that those two tags are attached to the same unique common resource, an email in our case. Thus, the name of the tag can be a locally defined string but the tag itself—including its semantic identity—is shared among Alice and Bob. The effect is that we allow users to rename tags locally, thus having access to unlimited aliases of their shared tags.

RESEARCH PLAN
We are interested in exploring how explicit support for sharing metadata in personal information management tasks can assist collaborative information seeking among close collaborators. Specifically, we seek to understand if our design reduces the time taken to locate information within one's information archives by such semi-automated human-dictated tagging.

Prototype Design
To explore our research questions, we are building a prototype based on Apple Mail[3]. This will allow users to perform tagging actions as part of their regular workflow, without needing to switch to an external application. For those users who do not use Apple Mail, we plan to create a second prototype, an out-of-process IMAP (Internet Mail Access Protocol) client that will let users filter their email accessed live from any IMAP server. Users can set specific per-collaborator per-tag privacy preferences. The tags will then be shared with other recipients of the same message (or the sender herself) if so allowed, and no others. It must be stressed that since neither message subjects nor content is sent to the tagging server, this poses minimal risk of inadvertent disclosure of confidential information to unauthorized parties.

Study Design
We plan to conduct a diary study combined with spot interviews of several participants using the prototypes to manage their email. Since email management practices evolve over a period of time and are intrinsically personal, it is important that any measures be taken after learning effects have been accounted for – at least several months for email, similar to [1]. In addition, we plan to instrument the tagging software to collect the following metrics to understand typical usage patterns: number of messages received, number of

tags suggested by collaborators, number of suggestions accepted, number of messages untagged after automated tagging, frequency of tagging and whether it was influenced by the presence of tag suggestions, % of messages left in inbox never tagged, and time required for re-finding tasks with automated tags applied. Timed micro-tasks will be administered during spot checks (e.g. *'please locate the last status update you received from Alice regarding your current project.'*) to evaluate the ability of users to locate information using the system.

REFERENCES
1. O. Bälter and C. Sidner. Bifrost inbox organizer: Giving users control over the inbox. In *NordiCHI '02: Proceedings of the Second Nordic Conference on Human-Computer Interaction*, pages 111–118, New York, NY, USA, 2002. ACM Press.

2. S. Bradshaw, M. Light, and D. Eichmann. (Bee)Dancing on the Boundary between PIM and GIM. In *Proceedings of the 2nd Invitational Workshop on Personal Information Management at SIGIR 2006.*, 2006.

3. R. G. Capra and M. Pérez-Quiñones. Re-finding found things: An exploratory study of how users re-find information. Technical report, Department of Computer Science, Virginia Tech, 2003.

4. H. H. Clark. *Using Language*. Cambridge University Press, New York, 1996.

5. H. H. Clark and S. E. Brennan. Grounding in communication. In L. B. Resnick, R. M. Levine, and S. D. Teasley, editors, *Perspectives on socially shared cognition*, pages 127–149. American Psychological Association Press, Washington DC, 1991.

6. H. H. Clark and D. Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

7. T. Erickson. From PIM to GIM: Personal Information Management in Group Contexts. *Commun. ACM*, 49(1):74–75, 2006.

8. D. Fisher, A. Brush, B. Hogan, M. Smith, and A. Jacobs. Using social metadata in email triage: Lessons from the field. *Human Interface and the Management of Information. Interacting in Information Environments*, pages 13–22, 2007.

9. C. Foley, P. Wilkins, and A. F. Smeaton. A collaborative video search system. In *CIVR 2009 - ACM International Conference on Image and Video Retrieval*, Santorini, Greece, 2009.

10. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Commun. ACM*, 30(11):964–971, 1987.

11. A. Garg, R. Battiti, and R. G. Cascella. "May I borrow Your Filter?" exchanging filters to combat spam in a community. In *AINA '06: Proceedings of the 20th*

---

[3]http://www.apple.com/macosx/features/mail.html

International Conference on Advanced Information Networking and Applications - Volume 2 (AINA'06), pages 489–493, Washington, DC, USA, 2006. IEEE Computer Society.

12. G. Golovchinsky, J. Pickens, and M. Back. A taxonomy of collaboration in online information seeking. In *Proceedings of the 1st Workshop on Collaborative Information Retrieval*, 2008.

13. J. Gwizdka. Email task management styles: The cleaners and the keepers. In *CHI '04: Extended Abstracts on Human Factors in Computing Systems*, pages 1235–1238, New York, NY, USA, 2004. ACM Press.

14. F. Hopfgartner, D. Vallet, M. Halvey, and J. Jose. Collaborative search trails for video search. In *Proceedings of the 1st Workshop on Collaborative Information Retrieval*, 2008.

15. E. C. Jaeger and I. H. Page. *A source-book of medical terms.* Charles C. Thomas, Springfield, Ill, 1953.

16. W. Jones. How is information personal? In *3rd Invitational Workshop on Personal Information Managementat CHI 2008: The Disappearing Desktop (PIM 2008)*, 2008.

17. H. Ma, R. Chandrasekar, C. Quirk, and A. Gupta. Page hunt: using human computation games to improve web search. In *HCOMP '09: Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 27–28, New York, NY, USA, 2009. ACM.

18. W. E. Mackay. More than just a communication system: diversity in the use of electronic mail. In *CSCW '88: Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, pages 344–353, New York, NY, USA, 1988. ACM Press.

19. D. Maier, A. Halevy, M. Bates, H. Bruce, B. Bederson, and M. Knox. Enhancements of personal information. In *Breakout Group Summary, PIM Workshop 2005*, 2005.

20. T. W. Malone. How do people organize their desks?: Implications for the design of office information systems. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 1(1):99–112, 1983.

21. C. Marlow, M. Naaman, D. Boyd, and M. Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia*, pages 31–40, New York, NY, USA, 2006. ACM Press.

22. M. R. Morris and E. Horvitz. SearchTogether: an interface for collaborative web search. In *UIST '07: Proceedings of the 20th annual ACM symposium on User interface software and technology*, pages 3–12, New York, NY, USA, 2007. ACM.

23. M. R. Morris and J. Teevan. Understanding groups' properties as a means of improving collaborative search systems. *CoRR*, abs/0908.0586, 2009.

24. J. Tang, T. Lau, J. Lin, and C. Drews. Consolidarity: Exploring patterns of social commonality among consolidated file storage. In *HCI Consortium Meeting, February 2006*. HCI Consortium, 2006.

25. J. Teevan, C. Alvarado, M. S. Ackerman, and D. Karger. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *CHI '04: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 415–422, New York, NY, USA, 2004. ACM Press.

26. L. von Ahn and L. Dabbish. Labeling images with a computer game. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA, 2004. ACM.

27. L. von Ahn, R. Liu, and M. Blum. Peekaboom: a game for locating objects in images. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 55–64, New York, NY, USA, 2006. ACM.

28. L. von Ahn, B. Maurer, C. McMillen, D. Abraham, and M. Blum. reCAPTCHA: Human-Based Character Recognition via Web Security Measures. *Science*, 321(5895):1465–1468, 2008.

29. S. Whittaker, S. E. Brennan, and H. H. Clark. Co-ordinating activity: an analysis of interaction in computer-supported co-operative work. In *CHI '91: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 361–367, New York, NY, USA, 1991. ACM.

30. S. Whittaker and C. Sidner. Email overload: exploring personal information management of email. In *CHI '96: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 276–283, New York, NY, USA, 1996. ACM Press.