A Position Paper on 'Living Laboratories': Rethinking Ecological Designs and Experimentation in Human-Computer Interaction

Ed H. Chi

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

Abstract. HCI have long moved beyond the evaluation setting of a single user sitting in front of a single desktop computer, yet many of our fundamentally held viewpoints about evaluation continues to be ruled by outdated biases derived from this legacy. We need to engage with real users in 'Living Laboratories', in which researchers either adopt or create functioning systems that are used in real settings. These new experimental platforms will greatly enable researchers to conduct evaluations that span many users, places, time, location, and social factors in ways that are unimaginable before.

Keywords: HCI, Evaluation, Ecological Design, Living Laboratories, Methodology, Web Services.

1 Introduction

Looking back on the history of Human-Computer Interaction as a field, we see fundamental contributions mainly from two groups of researchers: (1) computing scientists interested in how technology would change the way we all interact with information, and (2) psychologists (especially cognitive psychologists) interested in the implications of those changes. This created a combustible environment for great research, because the computing scientists wanted to create great and interesting tools but did not have a great way to measure its impact, yet many classically trained psychologists were looking beyond classic research in the brain and the understanding of human cognition. This resulted in an area called “Human Information-Processing”, which closely coupled with the growth of cognitive psychology, human factors, and human engineering [1, 11].

One enduring core value in Human-Computer Interaction (HCI) research has been the development of technologies that augment human intelligence. This mission originates with V. Bush, Licklider, and Engelbart, who inspired many researchers such as Alan Kay at PARC in the development of the personal computer and the graphical user interface. Together, both groups of researchers were excited by the possibilities of the computing machinery in producing systems that augmented human intellect [5], which was a possibility that was deeply intriguing to researchers that may have been

J.A. Jacko (Ed.): Human-Computer Interaction, Part I, HCII 2009, LNCS 5610, pp. 597–605, 2009.

slightly disillusioned with artificial intelligence research but yet believed computers were great tools for modeling and understanding human cognition. The aim of augmented human cognition has remained a core value for Human-Computer Interaction research.

With this aim, during the formation of the field, the need to establish HCI as a science had pushed us to adopt methods from psychology, both because it was convenient as well as the methods fit the needs. HCI field's rise paralleled the rise in the notion of personal computing---the idea that each person would have one computer at her command. Systems were evolving from many users using a single system to a single user multi-tasking with her own desktop computer. The costs of these systems forced researchers to think about how users would most productively accomplish knowledge work. The metaphor of the desktop, files, windows, and the graphical icons on bitmapped displays arrived naturally. The study of how users would respond to icons flashing on the screen, how users would move a pointing device like the mouse [2] to move a file from one location to the next paralleled some of the psychological experiments on stimulus and human response that psychologists were already routinely measuring. Fitts' law [1, 2], models of human memory [7], cognitive and behavioral modeling methods like GOMS [1] enabled HCI researchers and practitioners to model a single user interacting with a single computer.

2 Outdated Evaluative Assumptions

Of course, the world has changed. Trends in social computing as well as ubiquitous computing had pushed us to consider research methodologies that are very different from the past. In many cases, we can no longer assume:

Only a single display. Users will pay attention to only one display and one computer. Much of fundamental HCI research methodology assumes the singular occupation of the user is the display in front of them. Of course, this is no longer true. Not only do many users already use multiple displays, they also use tiny displays on cell phones and iPods and peripheral displays. Matthews et al. studied the use of peripheral displays, focusing particularly on glance-ability, for example. Traditional HCI and psychological experiments typically force users to attend to only one display at a time, often neglecting the purpose of peripheral display designs.

Only knowledge work. Users are performing the task as part of some knowledge work. The problem with this assumption is that non-information oriented work, such as entertainment applications, social networking systems, are often done without explicit goals in mind. With the rise of Web2.0 applications and systems, users are often on social systems to kill time, learn the current status of friends, and to serendipitously discover what might capture their interests.

Isolated worker. Users performing some task by themselves. Much of knowledge work turn out to be quite collaborative, perhaps more so than first imagined. Traditional view of HCI assumed the construction of a single report by a single individual that is needed by a hierarchically organized firm. Generally speaking, we have come to view such assumption with contempt. Information work, especially work done by highly paid analysts, is highly collaborative. Only the highly automated tasks that are routine and mundane are done in relative isolation. Information workers excel at exception handling, which often require the collaboration of many departments in different parts of the organizational chart.

Stationary worker. User location placement is stationary, and the computing device is stationary. A mega-trend in information work is the speed and mobility in which work is done. Workers are geographically dispersed, making collaboration across geographical boundaries and time-zone critical. As part of this trend, work is often done on the move, in the air while disconnected. Moreover, situation awareness is often accomplished via email clients such as Blackberries and iPhones. Many estimates now suggest that already more people access the internet on their mobile phone than on desktop computers. This certainly has been the trend in Japan, a bellwether of mobile information needs.

Task duration is short. Users are engaged with applications in time scales measures in seconds and minutes. While information work can be divided and be composed of many slices of smaller chunks of subgoals that can be analyzed separately, we now realize that many user needs and work goals stretch over for long period of time. User interests in topics as diverse as from news on the latest technological gadgets to snow reports for snowboarding need to be supported over periods of days, weeks, months and even years. User engagement with web applications are often measured in much longer periods of time as compared to more traditional psychological experiments that geared toward understanding of hand-eye coordination in single desktop application performance. For example, Rowan and Mynatt studied peripheral family portraits in the digital home over a year-long period and discovered that behavior changed with the seasons [14].

The above discussion point to how, as a field, HCI researchers have slowly broken out of the mold in which we were constrained. Increasingly, evaluations are often done in situations in which there are just too many uncontrolled conditions and variables. Artificially created environments such as in-lab studies are only capable of telling us behaviors in constrained situations. In order to understand how users behave in varied time and place, contexts and other situations, we need to systematically re-evaluate our research methodologies.

3 Re-thinking Evaluations

Fundamentally, traditional HCI research is busting the seams in two different ways: (1) ubiquitous computing research is challenging the notion of personal computing in front of a desktop, looking at computation that is embedded in the environment as well as computation done with ever powerful devices that can be taken while mobile [3, 4]; (2) social computing research that is simultaneously challenging the notion of computing systems designed for the individual, instead of for a group or community [6, 12].

Both trends have required re-thinking our evaluation methodologies. Traditional CSCW research have already drawn on qualitative methodologies from social scientists, including field observations and interviews, diary studies, survey methods, as well as focus groups and direct participation. Ubicomp, on the other hand, have used a mixture of methods, but have more readily examined actual deployments with real users in the field.

In either case, it may be time for us to fundamentally re-think how HCI researchers ought to perform evaluations, as well as the goal of the evaluations.

Since, increasingly, HCI systems are not designed for a single person, but for a whole group, we need research that not just augment human intelligence, but also group intelligence and social intelligence. Indeed, a natural extension of research in augmenting human intellect is the development of technologies that augment social intelligence, lead by research in the Social Web and Web2.0 movements. Traditional CSCW research has already studied the needs of coordination for a group and to some extent a community of practice. Many researchers are now conducting research in a social context, in which factors are less easy to isolate and control in the lab. Some research in the past might have treated variations in social contexts as part of the noise of the overall experiment, but this is clearly unsatisfactory since larger subject pools are necessary to overcome the loss in the power of the experiment. Moreover, we now know that many social factors follow distributions that are not normally distributed, making the prediction of individual factors in greatly varying social situations difficult, if not impossible.

Since users now interact with computing systems in varied ubiquitous contexts, ecological validity is often much more important than studying factors in isolation. In ubicomp applications, for example, productivity measurements are often not the only metrics that are important. For example, adoption of mobile applications is now often cited as evidence of the usefulness of an application.

One might argue that if using an application results in no productivity increase then the fact there is adoption of the application is irrelevant. However, this view is short sighted, because the opposite is also true: If there is productivity increase from using the application, but there is no adoption (perhaps due to ease of use issues, for example), then it is also unclear what benefit the application will ultimately bring. Obviously, the best situation is to have both productivity improvements as well as real adoption. However, research resource constraints often conspire against us to achieve both. Interestingly, academic research often tend to focus on the former rather than the latter, increasing the perceived gulf between academics' ivory tower and the trenches of the practitioners.

An example that illustrates this gulf is the studies around color copiers and printers. It has been circulated here at PARC that researchers had studied the need for color output from copiers and printers, and had concluded that there was either negligible increase or no productivity increase from using color. Cost and benefit analysis showed that black-and-white copiers were often just as good and more economical than color copiers in the majority of the cases. While it is unclear whether the studies took into account of increase use of color in various media might possibly drive future demand and utility of color systems, what is clear now is that the adoption of color copiers and printers would occur independent of productivity studies. If what matters in the industry are the adoption of technology, while academic research remains focused on measurements of productivity, we will never bring the two communities together and technology transfer will forever remain challenging.

4 Evaluations Using 'Living Laboratories'

The Augmented Social Cognition group have been a proponent of the idea of 'Living Labratory' within PARC[1]. The idea is that in order to bridge the gulf between academic models of science and practical research, we need to conduct research within living laboratories. Many of these living laboratories are real platforms and services that researchers would build and maintain, and just like Google Labs or beta software, would remain somewhat unreliable and experimental, but yet useful and real. The idea is to engage real users in ecological valid situations, while gathering data and building models of social behavior.

Looking at two different dimensions in which HCI researchers could conduct evaluations, one dimension is whether the system is under the control of the researcher or not. Typically, computing scientists build systems and want them evaluated for effectiveness. The other dimension is whether the study is conducted in the laboratory or in the wild. These two dimensions interact to form four different ways of conducting evaluations:

(1) Building a system, and studying it in the laboratory. This is the most traditional approach in HCI research and the one that is typically favored by CHI conference paper reviewers. The problem with this approach is that it is (1) extremely time-consuming, and (2) experiments are not always ecologically valid. As mentioned before, it is extremely difficult, if not impossible, to design experiments for many social and mobile applications that are ecologically valid in the laboratory.

(2) Not building a system (but adopt one), and still study it in the laboratory. For example, this is possible by taking existing systems, such as Microsoft Word and iWorks Pages and comparing the features of these two systems.

(3) Adopting an existing system, and studying it in the wild. The advantage here is to study real applications that are being used in ecologically valid situations. The disadvantage is that findings are often not comparable, since factors are harder to isolate. On the other hand, the advantages are that real findings can be immediately applied to the live system. Impact of the research is real, since adoption issues are already removed. As illustrated below, we have studied Wikipedia usage in detail using this method.

(4) Building a system, releasing it, and studying it in the wild. A well-publicized use of this approach is Google's A/B testing approach[2]. According to Google, A/B testing allowed them to finely tune the Search Engine Result Pages (SERPs). Some details about this kind of A/B online experiments has been documented [8]. For example, how many search results should the page contain was studied carefully by varying the number between a great number of users. Because the subject pool is large, Google can say with some certainty which design is better on their running system. A major disadvantage of this approach is the effort and resource requirement it takes to study such systems. However, for economically interesting applications

[1] http://asc-parc.blogspot.com/2008/11/living-laboratories-rethinking.html

[2] http://news.cnet.com/8301-10784_3-9954972-7.html

such as Web search engines, the tight integration between system and usage actually shorten the time to innovate between product versions.

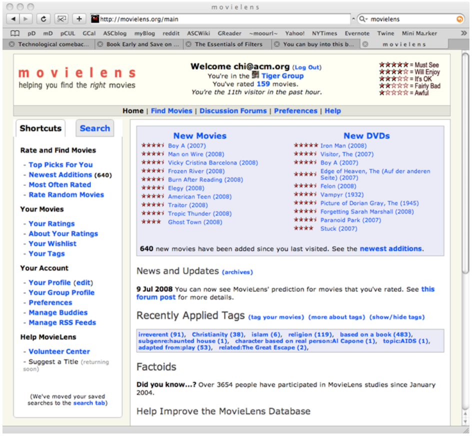
Of these variations, (3) and (4) are what we consider to be 'Living Laboratory' studies.

5 Examples of Living Laboratory Style Research

Here we will illustrate how to conduct Living Laboratory studies with some examples.

GroupLens and MovieLens. First, an example of building a real system, releasing it, and studying it in the wild was the seminal work of the GroupLens [9] research group at University of Minnesota. GroupLens was first created to deal with information overload, particularly the high amount of traffic in Usenet news. In this way, GroupLens was hoping to adopt an existing community and system, and augment it with some technology and studying how the technology performs in the wild. The technology in question was collaborative filtering. The idea at the time was related to user profiling. Users expressing interest in the same items must be somewhat similar and can form a virtual neighborhood. Therefore, we can recommend to them items that their neighbors are interested in. The research group was somewhat successful in doing this, as enough users on Usenet news adopted the technology and provided feedback on the system.

Later, the research group built a movie recommendation site on the Web that used similar collaborative filtering algorithms called MovieLens [10]. The website retained a community of about 6000 users that became an ecosystem in itself. Someone volunteered to keep the movie database up to date, and some participated in discussions about features the recommendation system should have. Later research on specific recommendation algorithms often split users into groups temporarily, where one group might receive one treatment, while the other would receive another treatment. The results are then compared to see how the two groups differed, including whether they evolved different group behaviors.


movielens
http://movielens.org/main
movielens
pD mD pCUL GCal ASCblog myBlog reddit ASCWiki GReader ~moourl~ Yahoo! NYTimes Evernote Twine Mini Ma.rker
Technological comebac...
Book Early and Save on ...
The Essentials of Filters
You can buy into this b...
movielens

movielens
helping you find the *right* movies

Welcome chi@acm.org (Log Out)
You're in the Tiger Group
You've rated 159 movies.
You're the 11th visitor in the past hour.

★★★★★ = Must See
★★★★☆ = Will Enjoy
★★★☆☆ = It's OK
★★☆☆☆ = Fairly Bad
★☆☆☆☆ = Awful

Home | Find Movies | Discussion Forums | Preferences | Help

Shortcuts Search

Rate and Find Movies

- Top Picks For You
- Newest Additions (640)
- Most Often Rated
- Rate Random Movies

Your Movies

- Your Ratings
- About Your Ratings
- Your Wishlist
- Your Tags

Your Account

- Your Profile (edit)
- Your Group Profile
- Preferences
- Manage Buddies
- Manage RSS Feeds

Help MovieLens

- Volunteer Center
- Suggest a Title (returning soon)

(We've moved your saved searches to the search tab)

New Movies

★★★★½ Boy A (2007)
★★★★½ Man on Wire (2008)
★★★★½ Vicky Cristina Barcelona (2008)
★★★★½ Frozen River (2008)
★★★★½ Burn After Reading (2008)
★★★★½ Elegy (2008)
★★★★½ American Teen (2008)
★★★★½ Traitor (2008)
★★★★½ Tropic Thunder (2008)
★★★★ Ghost Town (2008)

New DVDs

★★★★★ Iron Man (2008)
★★★★½ Visitor, The (2007)
★★★★½ Boy A (2007)
★★★★½ Edge of Heaven, The (Auf der anderen Seite) (2007)
★★★★½ Felon (2008)
★★★★½ Vampyr (1932)
★★★★½ Picture of Dorian Gray, The (1945)
★★★★½ Forgetting Sarah Marshall (2008)
★★★★½ Paranoid Park (2007)
★★★★ Stuck (2007)

640 new movies have been added since you last visited. See the newest additions.

News and Updates (archives)

9 Jul 2008 You can now see MovieLens' prediction for movies that you've rated. See this forum post for more details.

Recently Applied Tags (tag your movies) (more about tags) (show/hide tags)

irreverent (91), Christianity (38), islam (6), religion (119), based on a book (483), subgenre:haunted house (1), character based on real person:Al Capone (1), topic:AIDS (1), adapted from:play (53), related:The Great Escape (2),

Factoids

Did you know...? Over 3654 people have participated in MovieLens studies since January 2004.

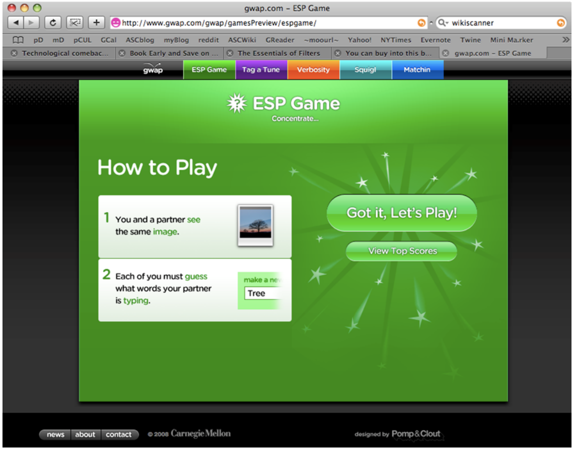
Help Improve the MovieLens Database

Fig. 1. Movielens system is an academic project with a live community

Games with a Purpose (gwap.com). Luis von Ahn's work on ESP games has evolved in a highly intriguing site called Games with a Purpose (gwap.com). On this site, users can engage in mini-games that are fun in themselves, but also the games end up collecting data that is useful in some other way. One well-known example is the image labeler, in which two users (without other communication means) must agree on the same keyword to receive points. The objective is to agree on the labeling in as many images as possible in a given timeframe.

Here the objective is to engage real users in realistic contexts, in which the goal was to entertain the user and to gather behavioral data that tell us something about the images. One can now analyze word choices over many data points, collective action (including any attempts at cheating), as well as longitudinal issues like number of repeat visits, the diversity of users, or viralness of the game.

Engagement measures, such as stickiness, can be directly measured.


gwap.com – ESP Game
http://www.gwap.com/gwap/gamesPreview/espgame/
wikiscanner
pD mD pCUL GCal ASCblog myBlog reddit ASCWiki GReader ~moourl~ Yahoo! NYTimes Evernote Twine Mini Ma.rker
Technological comebac... Book Early and Save on ... The Essentials of Filters You can buy into this b... gwap.com – ESP Game
gwap
ESP Game
Tag a Tune
Verbosity
Squigl
Matchin
ESP Game
Concentrate...
How to Play
make a ne
Tree
Got it, Let's Play!
View Top Scores
news
about
contact
© 2008 Carnegie Mellon
designed by Pomp&Clout

1 You and a partner see the same image.

2 Each of you must guess what words your partner is typing.

Fig. 2. The Games With A Purpose (gwap.com) website engages real users with games, while having them accomplish some task that is useful for research

WikiScanner / WikiDashboard over Wikipedia. One realistic approach is to adopt an existing community and system, and create mashup applications that augment the original system with some new capability and studying its effects. For example, Wikis are collaborative systems in which virtually anyone can edit anything. Although wikis have become highly popular in many domains, their mutable nature often leads them to be distrusted as a reliable source of information.

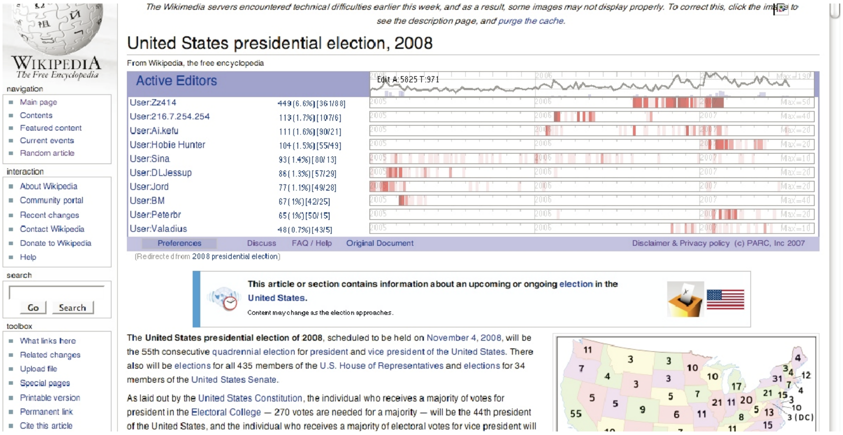
For example, Virgil Griffith took open source data from Wikipedia and enabled people to discover the possible identities of Wikipedia editors by cross-referencing the IP address with institution names[3].

Our own research on social transparency also took this approach. We downloaded a copy of all of the edits on Wikipedia and tabulated the editing statistics for all articles and all users[4]. This enabled us to create a visualization the editing patterns for

[3] http://wikiscanner.virgil.gr/

[4] http://wikidashboard.parc.com/

each article and each user [13]. WikiDashboard has received tens of thousands of visits from Wikipedia users. We know also that both systems were discussed extensively in the Wikipedia community.



The Wikimedia servers encountered technical difficulties earlier this week, and as a result, some images may not display properly. To correct this, click the image to see the description page, and purge the cache.

United States presidential election, 2008

From Wikipedia, the free encyclopedia

Active Editors		Edit A:5825 T:971 Max=190
User:Zz414	449 (6.6%) [361/88]	Max=50
User:216.7.254.254	113 (1.7%) [107/6]	Max=40
User:Ai.kefu	111 (1.6%) [90/21]	Max=20
User:Hobie Hunter	104 (1.5%) [55/49]	Max=20
User:Sina	93 (1.4%) [80/13]	Max=10
User:DLJessup	86 (1.3%) [57/29]	Max=20
User:Jord	77 (1.1%) [49/28]	Max=20
User:BM	67 (1%) [42/25]	Max=40
User:Peterbr	65 (1%) [50/15]	Max=20
User:Valadius	48 (0.7%) [43/5]	Max=10

(Redirected from 2008 presidential election)

This article or section contains information about an upcoming or ongoing election in the United States.

Content may change as the election approaches.

The **United States presidential election of 2008**, scheduled to be held on November 4, 2008, will be the 55th consecutive quadrennial election for president and vice president of the United States. There also will be elections for all 435 members of the U.S. House of Representatives and elections for 34 members of the United States Senate.

As laid out by the United States Constitution, the individual who receives a majority of votes for president in the Electoral College — 270 votes are needed for a majority — will be the 44th president of the United States, and the individual who receives a majority of electoral votes for vice president will

Fig. 3. An example page from WikiDashboard [13] project, which inserts a visualization of the social dynamics and edit patterns for every Wikipedia page

6 Conclusion

HCI research have greatly benefitted from borrowing evaluation methods that were fine-tuned in other fields, especially the behavioral sciences. Evaluation methods are inseparable from the kinds of science and models that can be build in a field. HCI have long moved beyond the evaluation of a single user sitting in front of a single desktop computer, yet many of our fundamentally held viewpoints about evaluation continues to be ruled by outdated biases derived from this legacy. In this position paper, we have argued that traditional views of human performance in systems have long been only focused on productivity. It is time for us to break out of these long-held views, and look at evaluations in more holistic ways.

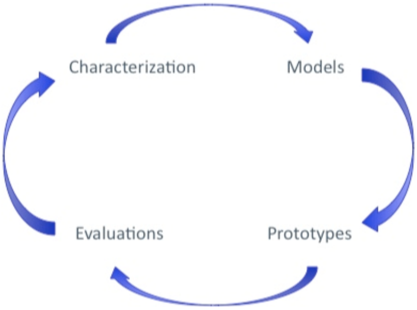


Fig. 4. A way to think about the role of Living Laboratory prototypes in scientific research

One way to do this is to engage with real users in 'Living Laboratories', in which researchers either adopt or create real useful systems that are used in real settings that are ecologically valid. This enables a tight loop between characterization of behavior, models of the users and system, prototype, and experimentation. The new Social Web platform is enabling researchers to build systems with amazing speed, enabling the whole loop to be completed within much shorter amounts of time than the past. Similar experimentation platforms for mobile computing is just becoming reachable, with iPhone and Google's Andriod leading the charge. These platforms will greatly enable Living Laboratory researchers to conduct evaluations that span many users, places, time, location, and social factors in ways that are unimaginable before.

Acknowledgments. We thank PARC's Augmented Social Cognition team and the HCIC workshop for many helpful discussions on this position paper.

References

1. Card, S., Moran, T.P., Newell, A.: The Psychology of Human Computer Interaction. Lawrence Erlbaum Associates, Mahwah (1983)
2. Card, S.K., English, W.K., Burr, B.J.: Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT. Ergonomics 21(8), 601–613 (1978)
3. Carter, S., Mankoff, J., Klemmer, S., Matthews, T.: Exiting the cleanroom: On ecological validity and ubiquitous computing. HCI Journal (2008)
4. Chi, E.H.: Introducing Wearable Force Sensors in Martial Arts. IEEE Pervasive Computing 4(3), 47–53 (2005)
5. Engelbart, D.C.: Augmenting Human Intellect: A Conceptual Framework. Summary Report AFOSR-3223 under Contract AF 49(638)–1024, SRI Project 3578 for Air Force Office of Scientific Research, Stanford Research Institute, Menlo Park, CA (1962)
6. Grudin, J.: Groupware and social dynamics: Eight challenges for developers. Communications of the ACM 37(1), 92–105 (1994)
7. Jones, W.P.: On the Applied Use of Human Memory Models: The Memory Extender Personal Filing System. International Journal of Man-Machine Studies 25(2), 191–228 (1986)
8. Kohavi, R., Longbotham, R.: Online Experiments: Lessons Learned. Computer 40(9), 103–105 (2007), doi:10.1109/MC.2007.328
9. Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R., Riedl, J.: GroupLens: Applying Collaborative Filtering to Usenet News in special section: recommendation systems. Communications of the ACM 40(3), 77–87 (1997)
10. Riedl, J., Konstan, J.: Word of Mouse: The Marketing Power of Collaborative Filtering. Warner Books, New York (2002)
11. Sears, A., Jacko, J.A.: The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications. CRC Press, Boca Raton (2008)
12. Shneiderman, B.: Science 2.0. Science 319(5868), 1349–1350 (2008)
13. Suh, B., Chi, E.H., Kittur, A., Pendleton, B.A.: Lifting the Veil: Improving Accountability and Social Transparency in Wikipedia with WikiDashboard. In: Proceedings of the ACM Conference on Human-factors in Computing Systems (CHI 2008), Florence, Italy, pp. 1037–1040. ACM Press, New York (2008)
14. Rowan, J., Mynatt, E.D.: Digital family portrait field trial: Support for aging in place. In: Proc. of CHI 2005 Conference on Human Factors in Computing Systems, pp. 521–530. ACM, New York (2005)