A Survey on Web Use: How People Access, Consume, Keep, and Organize Web Content

Seungyon Claire Lee, Eamonn O'Brien-Strain, Jerry Liu, Qian Lin

Keyword(s):

Content repository; web content; current practice; user experience

Abstract:

We present the results from a preliminary survey concerning how people access, consume, keep, and organize seven types of web content. From the results of the survey, we highlight design suggestion for Tangible Web (TW), our cloud content repository system that enables users to clip, save, format, and organize web content through active discovery and passive delivery.

A Survey on Web Use: How People Access, Consume, Keep, and Organize Web Content

Seungyon Claire Lee
Hewlett Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304 USA
claire.sylee@hp.com

Eamonn O'Brien-Strain
Hewlett Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304 USA
eamonn.obrien-strain@hp.com

Jerry Liu
Hewlett Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304 USA
jerry.liu@hp.com

Qian Lin
Hewlett Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304 USA
qian.lin@hp.com

Abstract

We present the results from a preliminary survey concerning how people access, consume, keep, and organize seven types of web content. From the results of the survey, we highlight design suggestion for Tangible Web (TW), our cloud content repository system that enables users to clip, save, format, and organize web content through active discovery and passive delivery.

Keywords

Content repository, web content, current practice
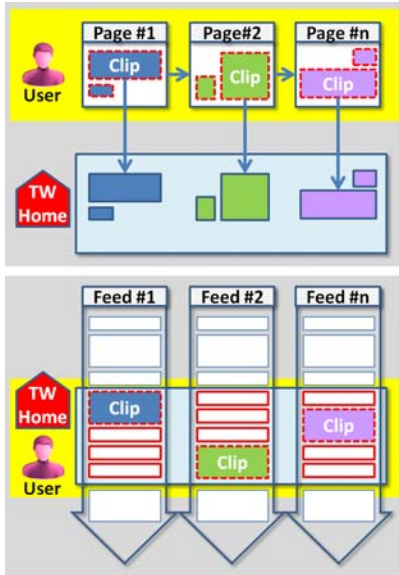
ACM Classification Keywords

K8.3. Personal computing: Management/maintenance

General Terms

Design.

Introduction

The history of graphical user interface (GUI) is tightly coupled with the continuous evolution of relevant user experience (UX). Even though they may look obvious and straightforward, interactions with many familiar metaphors within the GUI are the results from careful design iteration over the decades. In that sense, UX in current practice is efficient resource to reflect things

Figure 1. Two approaches to clip content: active discovery by which the user moves from page to page to clip resources (top) and passive delivery by which the user stays in TW home to read daily feed delivered from preregistered resources (bottom).

that can be accepted naturally when designing new systems. This paper presents the results from a preliminary survey that concerns lessons learned from current practice (how people access, consume, keep, and organize web content) to support the design of Tangible Web, a cloud content repository. Note that since cloud computing blurs the boundary of personal computing between local machine and web environment, the survey focuses on the web content while including some desktop metaphor as well.

Tangible Web: Cloud Content Repository

Tangible Web (TW) is a web content repository service that enables users to clip, save, format, and organize web content (image, text, video, URL) in a private account. The notion of clipping is inspired by the ephemeral nature of web content, which is frequently updated and may disappear for a given URL. Thus, similar to Zoetrope that concerns temporal change of web content by time [1], the clipped content in TW can be saved either as a live view that applies the updates of the pointer or a frozen view that captures the snapshot at the moment that the content is clipped. By using TW, users select multiple discrete areas in the web page as a single entity to clip so that any unwanted content on the web page can be disregarded (advertisement, page navigation object). The clipped content can be reformatted with other clipped entity contained in the TW repository. When a clipped entity is presented, either independently or organized with other clipped entities, the combined text streams reflow to generate appropriate page composition with the print-friendly formatting.

With respect to the method for accessing the content to clip, TW has two approaches depending on how the
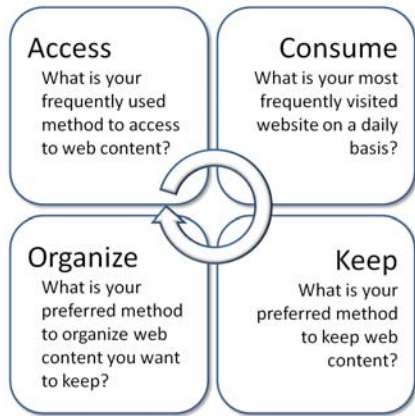
users contact the content of interest (Figure 1): active discovery and passive delivery. Active discovery refers to the situation where users are manually moving from page to page to search information to clip. In this case, TW is installed as an add-on of a web browser and is accessible any time while the user browses the web. EverNote (evernote.com) and Diigo (diigo.com) employ this approach.

On the other hand, passive delivery generates automatic content feed from preregistered resources such as users' existing social network service (SNS) or favorite websites. Examples of such approach include Google Reader *(google.com/reader*), FlipBoard (flipboard.com*)* and many personalized home pages (google.com/ig, my.yahoo.com*)*. In this approach, TW highlights how to integrate one-stop reading experience of daily updates with appropriate clipping capabilities.

Assuming that the nature of the content type (active discovery or passive delivery) may affect the relevant user behavior, we encountered several design questions while implementing TW (e.g., Will the content type and access method affect user behavior on clipping and organizing?). In search of answers for such questions, we perform a survey to learn about current practice on how people use web content on a daily basis. We expect that the lessons learned from such general behavior would assist us to predict requests from potential TW users and to find design strategy to fulfill these requests.

Study overview

The study specifies seven content types and four interaction types in the survey design.

Figure 2. Interaction cycle in four steps

Passive delivery	
<1>	Up-to-date news
<2>	Entertainment (video, music)
<3>	Personal content posted by others (blog, SNS)
<4>	Personal content posted by me (blog, SNS)
Active discovery	
<5>	Reference (knowledge, education)
<6>	Work related document (work, school)
<7>	Supporting document to assist real world activities (receipt, bill, coupon, banking info)

Table 1. Content type characteristics

Content type: Passive delivery vs. active discovery
The survey explores user behavior for interacting with seven content types that can rely on either passive delivery or active discovery (Table 1). Passive delivery content (<1> News, <2> Entertainment, <3> Content posted by others, <4> Content posted by me) may fit with the content that users want to know in a time-sensitive manner on a daily basis, whereas active discovery content (<5> References, <6> Work related document,-<7> Supporting document for real world activities) reflect content that might be less casual and would reflect users' moment-by-moment interests or urgent needs. Here, and throughout the paper, the content type will be described as a number bracket such as "<1>" for news content.

Interaction type: Four steps of behavior cycle
We cluster users' general behavior on the web as four steps within a cycle: access, consume, keep, and organize (Figure 2). In this cycle, behaviors on the web are somewhat reciprocal such that the result of one behavior often affects consecutive behavior (e.g., A bookmark that is used as a method for *access* is

created as a result of *organizing* a URL that is *kept*/saved on a browser during or after *consuming* the web content.). In the survey, behaviors in this four step cycle are integrated with relevant metaphors and systems such as bookmark and folder, metaphor in e-mail, file browser, and web browser. Researchers have studied matured UX in such systems with respect to the consumption of web media [7] and strategic management of emails by searching, tagging, and using folders [2, 4, 5, 8]. Unlike these studies that highlight a single behavior, our study focuses on the flow of the behavior cycle to understand overall trends in the interaction cycle.

Survey and participants
The survey includes four steps of behavior cycle (Figure 2) that are incorporated with seven content types (Table 1) and relevant metaphors. We recruited 435 volunteers to participant in the online survey (age 18-75, mean age = 35.54) who are college students or working professionals.

Result: Access

Behaviors and methods for access to the web indicate the moment that the user initiates the interaction with web content. Methods to access web content could be the results of organizing (e.g., Bookmark), which is another node in the interaction cycle. We integrate seven content types with four access methods (Type URL, search keyword, click bookmark, click link) to learn the relationship between content type and access method.

Only two methods were selected with content-specific behavior (bookmark and keyword search). The selection of access method is partially affected by the

content type. Keyword search is the most preferred method to access entertainment and reference content (<2, 5>), whereas bookmark is the most preferred method to access to SNS feeds (<3,4>), websites for work or school (<6>), and websites that provide supporting document to assist real world activity such as banking (<7>). For these content types, the preference on keyword search and bookmark is inversely proportional. However, no dominant method was observed for news access (<1>).

Question: Through what methods do you access the following types of content? (Select all applied each content type): Graph shows two main methods only
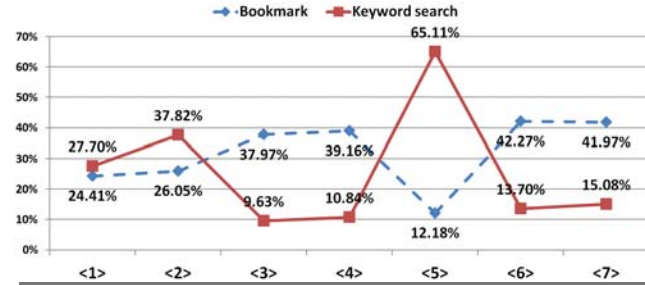

Figure 3. Access method by content type

Result: Consume

Questions in this interaction cycle explore content types by popularity to learn what people want to check from the web on a daily basis. We focus on integrating the popularity of website with the effect of popular website such as dominant leadership or long tail [2]. The notion of the long tail, which has been widely accepted to describe retailing strategy of selling (selling a large number of unique items each of which is sold with small

quantities), would help us to find the dominance or absence of the leading websites by content type.

Q: Excluding e-mail and search engines, what is the website that you access every day? (Open-ended)
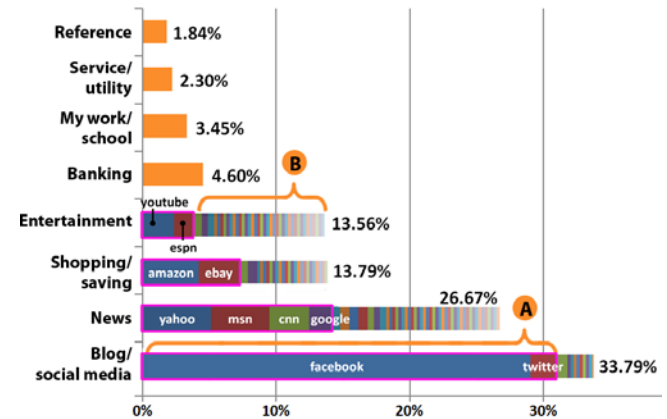

Figure 4. Daily visit of web content by popularity: (A) In SNS, two leaders attract most visitors (91.16%).; (B) In entertainment category, 39 non-popular websites attract visitors (72.88%) more than two leading websites.

Among 173 websites from 435 entries in eight categories (Figure 4), the ten most popular websites (Figure 5) and the rest 163 websites obtained 55.86% and 44.14% of votes, respectively. The top ten websites are included in top four dominant categories: two websites in SNS, four websites in news, two websites in shopping, and two websites in entertainment. Among these categories, the most and least significant effects of leading websites were found in SNS and entertainment categories, respectively. In SNS category, two leading websites attract large number of visitors (91.16%, Figure 4-A). On the other
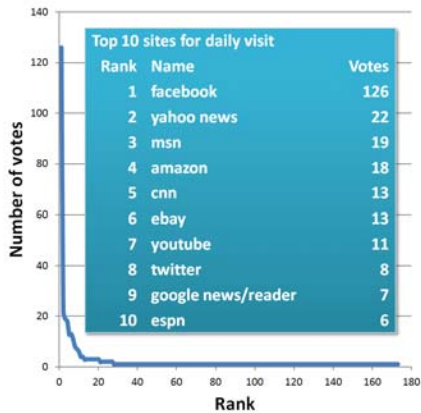

Figure 5. Ten most popular websites (small number of unique websites that have large number of visitors for each) and the rest 163 websites in a long tail (large number of unique websites that have small number of visitors for each)

hand, in entertainment category, large number of non-dominant websites are found (39 out of 41, 95.12%). Even though the number of visitors is small in each website, the sum is larger (72.88%, Figure 4-B) than the visitors of two leading websites.

People's daily visit on the ten most popular websites (55.86%) is slightly higher than the preference to visit the rest unique, yet non-popular websites (44.14%). The frequent visit on such diverse websites, which is especially observed in entertainment content, indicates the preference to consume content by personal interest rather than by common interest of others.

Result: Keep

With respect to the behavior on keeping web contents on a daily basis, which implies the moment that the public content becomes personal content, we are mostly concerned with how and why people select between saving and printing, what the relevant methods are, and how content type affects decision making.

Three personae were found in the question that asks about the preferred method among four choices (print, save, others, and not applicable) to keep content in seven content types (Figure 6): keeping web content very rarely (Group A), conditionally (Group B), and actively (Group C). Group A (14.05%) does not keep web content in any content type. Group B (38.10%) conditionally keeps web content. They keep at least one content type that relies on active discovery (<5, 6, 7>) while keep none of the content in passive delivery category (<1>-<4>). Group C, which shows the most active keeping behavior, keeps all types of given content.
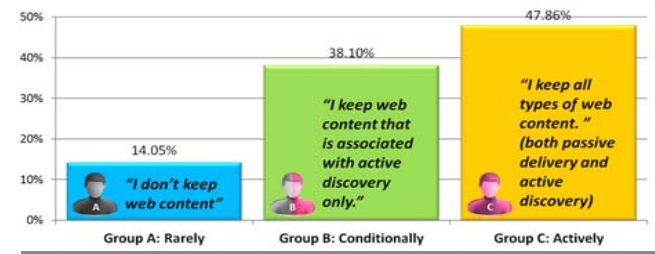

Figure 6. Three personae in keeping web content.

Since different personae were found, analyzing behavior to keep web content is performed with Group B and C. Group A result is not included in the analysis. In Figure 7 and 8, the result from Groups B and C are illustrated as line and bar graph, respectively, such that line graph is found only in content <5>, <6>, and <7>. Stars on each graph present the dominant method of which the preference is statistically significant after CHI-square (p<.05). Empty and filled stars stand for the result from Group B and C, respectively.

Q: Which method do you use to keep web content as yours? (Select one method for each content type.)
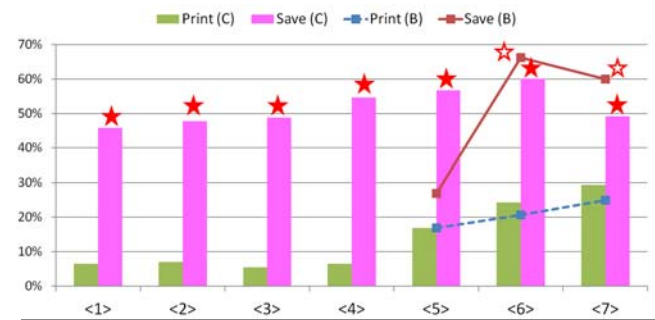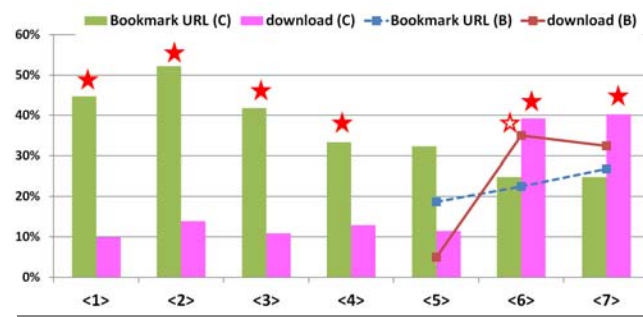

Figure 7. Keeping web content by content type and persona

Group B (Figure 7) prefers saving to printing for keeping web content, with statistical significance ($p<.05$) in work related document (<6>) and supporting document for daily activity such as shopping or banking (<7>). Group C prefers saving to printing ($p<.05$) to keep all content type.

To learn about preferred method to save web content, we provided five methods in another question: bookmark, downloading to the local machine, uploading to cloud storage, post to SNS, copy and paste to the local machine text editor and save. Among these methods, bookmark and download are mostly preferred. The preference for downloading work related document (<6>) to the local machine is discovered from Group B (Figure 8, $p<.05$). For all content types in passive delivery category (<1, 2, 3, 4>), Group C prefers bookmark ($p<.05$). However, for <6> and <7> in active discovery category, Group C prefers to download the content into the local machine ($p<.05$).

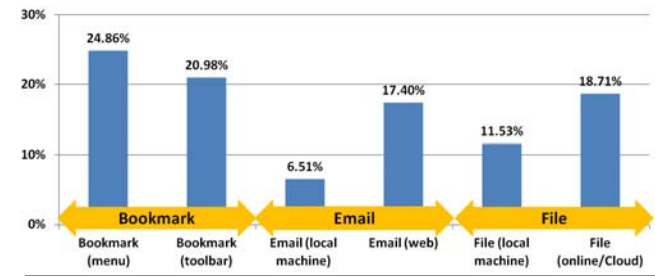*Q: Which method do you use to save web content?*

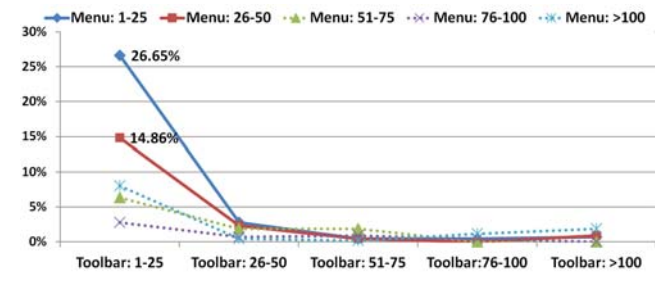**Figure 8.** Saving web content by content type and persona

Result: Organize

The behavior for organization is related with the preferences and strategies to manage personal information by using three methods: bookmark, email, and file browser. The preference for using bookmarks, email, and file browser are 45.84%, 23.91%, and 30.24%, respectively (Figure 9).

Q: How do you organize content saved from the web? (Select all applied)

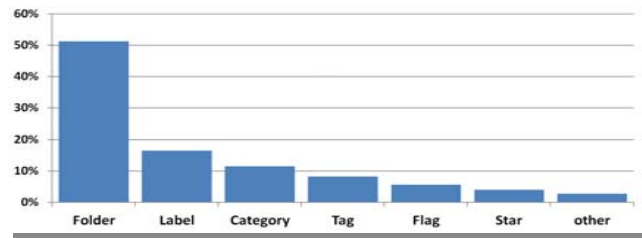**Figure 9.** Organizing content saved from the web

Q: How many bookmarks do you have in bookmark menu and toolbar in your default web browser?

Figure 10. Managing bookmark by using toolbar and menu

Q: How do you further organize/categorize to manage bookmark, emails, and files? (Select all applied)

Figure 11. Methods to organize content saved from the web

The use of bookmark toolbar significantly drops when the number of bookmarks exceeds 25 (Figure 10). This tendency is stronger when the number of bookmarks in the menu is less than 50 (solid lines in Figure 10). When the number of bookmarks in the toolbar is less than 25 (Toolbar:1-25 in Figure 10), people tend to have more items in the bookmark menu. However, this difference is not observed if the number of bookmarks on the toolbar exceeds 25. Among various metaphors that assist organizing, folder (51.27%), which probably has the longest history of use and is widely applied, is the most preferred metaphor (Figure 11).

	Access	Consume	Keep
<1>	Various	2nd popular content. Has decent leaders.	Bookmark
<2>	Search	4th popular content. Has weak leaders.	Bookmark
<3>	Bookmark	1st popular content. Has strong leaders.	Bookmark
<4>	Bookmark	1st popular content. Has strong leaders.	Bookmark
<5>	Search	Less popular content. No leaders.	Various
<6>	Bookmark	Less popular content. No leaders	Download
<7>	Bookmark	3rd popular content. Has decent leaders.	Download

Table 2. Summary

Design implication
Lessons learned from this survey (Table 2) contribute to building content-specific design strategies of TW, which aims to include contents by both passive delivery and active discovery (Figure 1).

Access: Bookmark vs. Keyword search
Bookmark is the most dominant method to access SNS, work related document, and supporting document (<3, 4, 6, 7>). Thus, TW can actively utilize bookmarks as a source of automatic or passive delivery. However, since not only those contents but also bookmarks could be highly private by the nature, TW should consider privacy issue carefully when augmenting daily interaction. On the other hand, keyword search is preferred to accessing entertainment and reference content (<2, 5>). Such content could be highly situational, hard to predict, and require manual operation. Thus, automating such activity through user analytics would be helpful in predicting user interest and profile. As Liu et al. developed a system to recommend personalized news content by analyzing click behavior [6], entertainment and reference content (<2>, <5>) can be recommended in TW by analyzing history of relevant activities: content clip; keyword search; and link click.

Consume: Find the niche for entertainment portal
Unlike SNS (<3, 4>), entertainment content (<2>) does not have a significant leader. Among 41 websites that constitute entertainment content, two top websites draw only 27.12% of traffic, which implies relatively weak hold on user consumption and long tail of the market share. Such a long tail (72.88% of traffic from 39 websites) refers to the need of entertainment portal that aggregates various, yet user-specific, requests.

Keep: Passive delivery vs. active discovery
Passive delivery content is saved by bookmarks. However, active discovery content, except references, is downloaded to a local machine, which is a more private space. Reference content (<5>) is saved by various methods with similar preference (bookmark, post to SNS, or download). From this result, we assume that TW should carefully detect such user preference. The fact that the news, entertainment, and SNS feeds (<1>-<4>) are mostly saved as bookmark reflects the characteristics of the content that is consumed, updated, and disposed frequently. Thus, TW may highlight the timely delivery and appropriate visibility of such content while concerning information overload. On the other hand, work related documents and supporting documents for the real world activities (<6>, <7>) may require advanced features saved in TW repository: ensuring secure management and clustering saved content with other clipped content of other application for further modification.

Organize: Integrating diverse concept with simplicity
In this step, even the content clipped from the public web is considered as a personal resource. We observe that the dominant preferences are for bookmark and folder metaphors. In TW, which will eventually blur the boundary among relevant applications, current practice to organize bookmarks on the web browser, files in the file browser on a local machine, and folders or other metaphors on the email client will be seamlessly aggregated. Thus, the challenges in designing an organizing assistant in TW would be the integration of diverse concepts with simplicity: how to optimize diverse organizing method to fulfill various need and context and how to provide concrete structure in a universal manner to convey general theme in the UX

design. Although not explored in this preliminary survey, examining the effects of content type with respect to the relationship between browsing history and organizing strategy in emails, bookmarks, and files would be a viable direction for further investigation.

Acknowledgement

References
[1]   Adar, E., Dontcheva, M., Fogarty, J. and Weld, D. S. 2008. Zoetrope: interacting with the ephemeral web. In *Proc. UIST 2008*. ACM Press (2008), 239-248.

[2]   Ames, M. and Naaman, M. Why we tag: motivations for annotation in mobile and online media. In *Proc.* CHI 2007. ACM Press (2007), 971-980.

[3]   Anderson, C. *The long tail: why the future of business is selling less of more*. New York: Hyperion, 2006.

[4]   Fisher, D., Brush, A. J., Gleave, E. and Smith, . A. Revisiting Whittaker & Sidner's "email overload" ten years later. In *Proc.* CSCW 2006. ACM Press (2006). 309-312.

[5]   Jones, W., Phuwanartnurak, A. J., Gill, R. and Bruce, H. Don't take my folders away!: organizing personal information to get things done. In Proc. CHI EA 2005. ACM Press, 1505-1508.

[6]   Liu, J., Dolan, P. and Pedersen, E.R. Personalized news recommendation based on click behavior. In *Proc.*IUI 2010. ACM Press (2010), 31-40.

[7]   Savolainen, R. Filtering and withdrawing: strategies for coping with information overload in everyday contexts. *J. Inf. Sci.* 33, 5 (October 2007), 611-621.

[8]   Whittaker, S. and Sidner, C. Email overload: exploring personal information management of email. In *Proc.* CHI 1996, ACM Press (1996). 276-283.