Towards Understanding Digital Information Discovery and Curation

Elena Voyloshnikova
elenavoy@uvic.ca

Dr. Margaret-Anne Storey
mstorey@uvic.ca

University of Victoria
Victoria, BC, Canada

Abstract

Everyday life revolves around the discovery and curation of digital information. People search the Web continuously, from quickly looking up the information needed to complete a task, to endlessly searching for inspiration and knowledge. A variety of studies have modelled information seeking strategies and characterized information seeking and curation activities on the Web. However, there is a lack of research on how existing Web applications support the discovery and management of information, especially concerning the motivations behind them and how different approaches can be compared.

In this paper, we present a study of information discovery tools and how they relate to the nature of information seeking. We propose a conceptual framework that deals with the opportunistic and purposeful aspects of how people discover and manage digital information. This framework can be used when designing new, evaluating, or updating existing Web applications to match expanded use case objectives.

1 Introduction

Today, people commonly use Web technologies to satisfy their information needs. For example, some people research their travel destinations using various online resources, some shoppers look for product characteristics to make purchasing decisions online, and some look up precise locations when needed. In order to accommodate diverse and evolving user needs, web applications continuously introduce new features and services empowering information discovery and curation.

Sometimes, Web users hope to find particular pieces of information, such as showtimes and phone numbers, to satisfy specific information needs. Other times, users lack well-articulated information needs, which results in their seeking behavior and knowledge gaps continuously changing. Often, people discover information online without even looking for it. The nature of information discovery can vary, and therefore, different types of it can require different tool support.

In addition, people perform information curation tasks to manage, preserve, and add value to collections of information. With rapidly increasing popularity of socially curated information spaces, it is important to understand how to enable curation activities when designing tools to support information discovery.

To close their knowledge gaps, people turn to various Web technologies ranging from specialized search tools to visual discovery applications. Several studies have been directed at exploring high-level Web tasks, including information seeking tasks [17, 11, 12, 18], deriving models of information seeking behaviors [5], and looking at methods of information curation [19]. However, more research is necessary to determine how different tools and their features provide fundamental support for information discovery and curation.

To enhance information seeking and curating experiences and support users' interactions, we extend existing research by (1) deriving factors that enable information discovery and curation and relating them within a framework, (2) validating the framework by studying and describing currently existing web applications, and (3) providing questions to ask to address concepts of the framework when evaluating and designing new applications. Specifically, our research goal is to gain understanding of how existing tools support digital information curation and discovery.

The remainder of this paper is organized as follows. Section 2 highlights some of the studies and technologies related to information curation and seeking tasks and behavior. Section 3 outlines the conceptual framework of factors and question that enable digital information discovery and support curation. In Section 4, we describe the methodology for validating the framework, and in Section 5 we report sample design strategies that address the concepts in today's Web applications. Section 6 outlines the findings and relates them back to the literature. In Section 7, we describe the final framework, followed by future work and conclusion in sections 7 and 8 respectively.

2 Related Work

Several researchers have studied various aspects of Web-based information discovery. To gain understanding of how currently existing web tools support information discovery and curation, we first study known characteristics of information-related web usage, including high-level web tasks, information seeking behavior, information curation, and modes of web use.

Kellar et al. [11] separated Web tasks into five categories: transactions, browsing, fact finding, information gathering, and other uncategorized tasks, with browsing, fact find, ing, and information gathering composing information seeking. Although the authors categorized information gathering as part of information seeking, it is in fact more closely related to digital curation [4, 19]. In their later work, Kellar et al. [12] added communication and maintenance as additional Web tasks.

Similarly to Kellar, Sellen et al. [18] identified six web tasks that are commonly performed by Web users: browsing, finding, housekeeping, information gathering, communicating, and transacting. Therefore, Kellar et al. and Sellen at al. both identified browsing, fact finding, and information gathering as information-related tasks that users perform online. Although this categorization is useful, it might not be enough to develop in depth support for information discovery and curation online.

People often engage in information seeking activities to close some knowledge gap that occurred as a result of not having enough information to perform some task [15]. Therefore, when providing tool support for various information discovery tasks, it is useful to consider the motivation behind these tasks, as it may differ. Morrison et al. [17] makes a distinction between methods of Web use and purposes. The authors derived a purpose taxonomy Web use, including three purposes or motivations: finding information, comparing or choosing to make a decision, and using the Web to find relevant information to gain understanding of some subject. Consequently, methods of finding information identified by Morrison et al. are collecting, finding, exploring, and monitoring. The differences between the two taxonomies suggest that different information seeking tasks may be performed to satisfy more than one information seeking purposes. Therefore, each purpose may require more than one task-supporting mechanisms.

A number of researchers have studies information seeking behavior [2, 3, 5, 6, 8, 7]. Ellis et al [6, 8, 7], proposed a model of information seeking characterized by six different patterns: starting, chaining, browsing, extracting, monitoring, and differentiating. Based on these patterns, Choo et al. [5], derived corresponding anticipated web moves. According to the authors, when users identify sources of interest, they usually identify which web sites can point to that information of interest. Chaining occurs when users navigate through links on those initial pages. When people browse, they scan top-level pages, headings, lists, and site maps. Differentiating takes place when people bookmark, print, copy and paste information, or choose earlier selected site. Monitoring occurs when users revisit web pages or receive updates from some earlier visited sites. Finally, extracting can occur when the user systematically searches the site to extract information of interest.

In 1986, Bates [3] proposed a model of four information seeking modes that consists of being aware, monitoring, browsing, and searching. Bates differentiated the modes based on the levels of at-

tention being active or passive, and information needs being directed or undirected. Thus, browsing can be characterized as undirected active information seeking because users do not know directly what information they are looking for, but they actively look for some information. Searching falls under active directed information seeking because the information need is clearly defined s the search is directed. Finally, monitoring and being aware are passive modes of information seeking although monitoring is directed, and being aware is undirected.

In 2002, Bates [2] extended her research with the notion of information farming. Information farming involves people collect and organize information for future use and revisitation. More commonly, it is referred to as digital curation, which is the notion of collecting and managing digital information for the purpose of adding value to the collection, and revisitation [4]. Wittaker [19] believes that in terms of Web use, significant shift is happening from information consumption to information curation, which means that people no longer just use the web to find and consume the information that they are interested in, but they also try to save and manage that information so that later it can be reaccessed and exploited.

Categorizing web usage into information seeking, digital curation, and other web tasks does not necessarily give full insight about how information-related tasks are performed. Lindley et al. [14] conducted a qualitative study involving 24 participants tracking their daily web usages in a form of a diary. As a result of this study, the researchers identified five distinct modes of web use: respite, orienting, opportunistic, purposeful, and lean-back. According to authors, people web browse opportunistically when they look for information related to some personal interest, long-term goal, or future ambition. Purposeful use occurs when the user knows for a fact what information she needs to acquire or what action online she needs to perform in order to continue or finish some other activity. Respite mode usually occurs when users are in the process of waiting for something or taking a break, and it serves as means for people to temporarily occupy themselves when high engagement with the content is not a requirement. Orienting mode usually occurs when people want to be updated on what has been happening in their environment. Examples of this mode are checking email at work or looking at the news and updates on a social networking site. Finally, lean-back mode of web use can be thought of as listening to the radio or watching TV. It usually involves watching videos online or browsing through other types of entertainment content.

Lindley's et al. primary motivations behind looking at use modes that occur when people browse the Internet was that traditional Web use studies and Web tasks discovered by other researchers could not reflect the depth of user's intentions online. Understanding characteristics of different modes can help designing models of Web interactions. For example, opportunistic use can have blur and continuously changing information need. People often cannot indicate completion of the web task, and they finish whenever they have been browsing Internet for too long, or they need to complete some other task of higher priority. Then, they often resume their opportunistic information seeking. Finally, opportunistic use is 'grasshopper-like' - users jump from one resource to another. From these factors, we can assume that to support such Web use, we would need to consider mechanisms for suggesting users information needs, and support revisitation and arbitrary navigation.

A number of researchers have studied how people search for and curate information, including information seeking tasks, information seeking behaviors, and information curation. However, there is lack of research on how currently existing technologies assist people in performing these tasks. With this overview of related work in mind, we proceed with building a framework of Web application design factors that enable information discovery and curation.

3 Conceptual Framework

Although information discovery and curation tasks elicit predominant portion of interactions within many Web applications today, there is lack of systematic guidelines for designing and evaluating applications that could afford these tasks. We hope to reduce this gap by developing a framework of design factors that enable digital information discovery and curation.

The framework (see Table 1) consists of two main categories of factors, discovery and curation, that are consequently broken down into subcategories. Each subcategory contains factors that de-

Table 1: Conceptual Framework.

Factors	Questions
Discovery	
Serendipitous discovery	
Arbitrary navigation	Does the application provide means for arbitrary navigation among resources?
Search-based navigation	Does the search feature help discover arbitrary resources related to the topic of interest?
Navigation within a directory	Do directory categories suggest and help navigating to resources related to the topic of interest?
Integration	Does the application support navigation to information sources beyond the application?
Visual preview	Do resources have visual previews?
Spatial arrangement	Are resources presented in a spatially meaningful way?
Fact discovery	
Search-based navigation	Does the search feature help discover the specific resource of interest?
Navigation within a directory	Do directory categories help narrow results to specific types of resources?
Uniform representation	Are resources presented in a uniform way?
Rediscovery	
History-based rediscovery	Does the application save and provide access to browsing history?
Bookmark-based rediscovery	Does the application support bookmark-based resource revisitation?
Search-based rediscovery	Is the search within application a reliable method for resource revisitation?
Subscription-based discovery	
Site subscription	Does the application allow subscriptions to news and updates?
User subscription	Does the application allow subscriptions to other users' activities?
Notifications	Does the application have notification mechanism(s)?
News stream	Can subscription updates be visible within the application?
Curation	
Management	
Categorization	Does the application support information categorization?
Preservation	
Internal Preservation	Does the application have bookmarking mechanism for preserving information within the application?
External Preservation	Does the application have bookmarking mechanism(s) for preserving information outside of the application?
Information enhancement	
Evaluation	Can resources be evaluated?
Annotation	Can resources be annotated?
Social curation	
Adding resources	Can resources be added to the publicly available pool of information from outside of the application?
Resharing resources	Can resources be publicly reshared within the application?

termine use case enablers and corresponding questions to ask when designing or evaluating an application. This section outlines main components of the framework.

3.1 Information Discovery

A number of researchers attempted classification of information seeking tasks and methods. In our framework, we built on existing research to derive corresponding design factors. Thus, the discovery category consists of serendipitous discovery, fact discovery, rediscovery, and channel-based discovery.

3.1.1 Serendipitous discovery

Serendipitous discovery refers to information discovery resulting from serendipitous browsing. Such discovery is characterized by underdefined, absent, or hidden information need, and it is usually involves browsing through diverse resources with varying content types [11]. The following is the key criteria that influence serendipitous information discovery:

- **Arbitrary navigation.** In order to browse through diverse information, an information discovery tool needs to provide a way to arbitrary navigate among resources adapting to continuously changing information needs and interests.
- **Search-based navigation.** Search-based navigation helps retrieve information sufficient for identifying an entry point of search [13]. In case of serendipitous discovery, since the information need is not well-articulated, the search feature should provide resources related to a broad topic of interest.
- **Navigation within a directory.** Similarly to search-based navigation, navigation within a directory should provide a way to narrow the results to those related to one topic. In addition, navigation with a dictionary can suggest topics of interest and help the user formulate an information need [13].
- **Integration.** To users with ambiguous information needs, one information portal might not provide access to all information of interest. If an information discovery application gives access to resources from various sources, the user should be able to navigate to original sources.
- **Visual preview.** Abrams et al. [1] identified link representation as one of the problems with traditional bookmarking. Analogously with browsing through a bookmark manager, identifying relevant information when browsing through links to diverse resources can be a challenging task. Visual preview should make it easier to evaluate relevance of resources.
- **Spatial arrangement.** Similarly to link representation, spatial visualization of numerous links is another problem that occurs when browsing through diverse content [1]. Therefore, meaningful spatial arrangement is important when displaying links to diverse resources.

3.1.2 Fact discovery

Fact discovery refers to information discovery resulting from looking for a specific piece of information. It is characterized by a well-defined information need and is easier to perform within systems that provide access to homogeneous types of information [11, 14]. Below is the list of factors that influence fact discovery:

- **Search-based navigation.** Since with fact discovery an information need is known, the goal of search-based navigation is to navigate to the resource of interest as directly as possible. Therefore search should limit resources to make it possible to find the one resource of interest.
- **Navigation within a directory.** Navigation within a directory is used to direct the user to relevant resources. In case of fact discovery, it should narrow the results to a specific type of resource so that further fact discovery is bounded by that type.
- **Uniform representation.** Uniform representation is a method of displaying diverse resources in a common way, with each resource having the same set of components. Such representation assures that each resource has the same set of facts associated with it, and therefore, the user can afford to have expectations about information that she can find when looking for a specific fact.

3.1.3 Rediscovery

Rediscovery refers to to information discovery resulting from revisiting previously discovered resources. The following is a list of factors that enable rediscovery:

- **History-based rediscovery.** A Web application needs to automatically backup browsing history in order to enable history-based rediscovery. History-based rediscovery can be useful when no bookmarking took place when the resource was discovered for the first time.
- **Bookmark-based rediscovery.** Bookmark-based rediscovery is one of the most common ways of information revisitation. The majority of Web browsers are equipped with bookmarking mechanisms. However, some modern Web applications provide integrated mechanisms for bookmarking and bookmark-based information rediscovery within application itself.
- **Search-based rediscovery.** Search-based rediscovery is not always a reliable way of refinding information. In information portals that provide access to fairly ambiguous information and that have information regularly repopulated and updated, search is usually designed around retrieving information related to some topic but not very specific. In order to revisit a resource, search must provide consistent results.

3.1.4 Subscription-based discovery

Subscription-based discovery is based on a combination of monitoring and awareness types of information seeking. It occurs when information is suggested to users based on their subscriptions. If users can actively look for updates, then an application affords monitoring. If users can receive notifications about updates, then an application facilitates awareness.

- **Site subscription.** Subscriptions to updates from a site help users follow the news. In order to support subscription-based discovery, an application must provide a subscription mechanism.
- **User subscription.** Similarly to site subscriptions, user subscriptions help following individual users and their activities. Such subscriptions help to further filter the content delivered to the user.
- **Notifications.** Notification mechanisms enable user awareness about the new content on the subscribed channel.
- **News stream.** Displaying news stream within application further promotes awareness and can serve as a monitoring mechanism.

3.2 Information Curation

By definition, digital information curation is the notion of managing, preserving, and adding value to collections of information. Thus, the curation category consists of information management, preservation, information enhancement, and social curation.

3.2.1 Management

Information management is one of the key elements of information curation. In the context of Web information management, the following factor plays the major role:

- **Categorization.** Resource categorization helps establish relationships between various resources. Allowing to customary categorize information can aid rediscovery, discovery in a socially curated space, as well add value to the resource.

3.2.2 Information Preservation

Information preservation, or information gathering, is a common Web tasks that is usually performed with an intent of information revisitation. However, in the case of opportunistic web use, information gathering is sometimes performed with just a goal of collecting information rather than revisiting it in the future. Traditional information preservation mechanism is bookmarking. Many Web applications apply variations of bookmarking mechanisms.

- **Internal Preservation.** Internal information preservation means bookmarking resources within application to be accessible within the same application. Such bookmarking facilitates information curation within the system.
- **External Preservation.** External preservation means bookmarking resources to be available within some other bookmarking system.

An application must facilitate integration with other applications in order to enable external preservation.

3.2.3 Information Enhancement

One of the important elements of digital curation is adding value to information overtime. In other words, enhancing already present information. Some forms of information enhancement are prevalent mainly in the context of social curation. Others can be performed in the context of personal curation.

- **Evaluation.** Evaluation methods vary from ratings to 'likes'. They usually take place in socially curated information systems. However, some of them contribute to personal reflection and information preservation.
- **Annotation.** Annotations are metadata attached to a resource. They can take have forms of comments, tags, descriptions, etc. Annotations make it easier to search for resources as well as to interpret them.

3.2.4 Social Curation

In addition to public resource sharing, social information curation encapsulates all of the elements of digital curation. Therefore, adding resources and resharing those resources are the two key factors that empower social curation.

- **Adding resources.** Adding resources does not only facilitate global Web information curation but also scales the information available through the system up providing more opportunities for information discovery.
- **Resharing resources.** Resharing currently existing within the system resources supports subscription-based information discovery since it supports channeling information that can be of interest to subscribed groups of users.

The following section describes the methodology for evaluating the conceptual framework and understanding how to address the elements of the framework when designing real world applications.

4 Methodology

The study presented in this paper has two primary goals. The first goal is to validate the conceptual framework of factors that enable digital information discovery and curation. The second goal of the study is to gain perspectives on how to address different elements of our framework when designing real-world applications. Therefore, our research questions are the following:

RQ1: How do existing Web applications support information discovery?
RQ2: How do existing Web applications support information curation?

Our methodology for studying existing Web applications is based on Yin's guidelines for designing a case study [20]. The motivation behind choosing a case study over other methods of qualitative research was based on our choice of research questions, which have an explanatory nature, lack of control over existing applications and their development, and having to focus on contemporary use of real-life web applications. According to Yin [20], case study would be an optimal research strategy given above characteristics of the subject matter.

To answer our research questions, we analyzed twenty different cases. For each case, we picked a Web application whose primary purpose is to be used for information discovery. We examined overall purpose of each application, its description as defined by the site itself, literature and documentation related to the case, if they were available, against the features that the application provides. For example, if an application provides bookmarking features, we would check if it is indeed intended to be used for information preservation.

To increase external validity of our study, we chose the cases based on replication logic [20]. Using replication logic in a case study design means carefully selecting each case so that it either predicts analogous results or predicts contrasting results but for anticipated reasons. Therefore, we used our conceptual framework (see Section 3) to predict if an application supports each of the information discovery and curation tasks based on the features that the application provides.

Consequently, our case study had a form of an iterative process of selecting cases, analyzing each case, and determining whether or not it meets our theoretical propositions. If one or more cases did not support the theory, then we modified the propositions and selected a new set of cases until the re-

sults of analyzing the case gave the anticipated results for all cases. We then grouped features into factors that enable information discovery and curation, and recorded corresponding questions to ask in order to evaluate an application. The resulting framework is depicted in Table 1. Limitations of our study are presented in Section 6.

5 Digital Information Discovery and Curation

Through examining twenty different Web applications whose main function is to help people discover information, we were able to build and validate the conceptual framework of factors and questions that facilitate digital information discovery and curation. In this section, we present our findings and answer the research questions, RQ1 and RQ2.

RQ1: How do existing Web applications support information discovery?

There is a large spectrum of usages within information discovery systems.

Applications that facilitate serendipitous information discovery often employ elaborate resource representation techniques. Social-bookmarking systems, such as Pinterest, Scoop.it!, and StumbleUpon, had indeed tried solving the two link representation problems indicated by Abrams [1], visualization and representation.

Search-based rediscovery can be a challenging task within some applications and a very simple in others.

RQ2: How do existing Web applications support information curation?

6 Limitations and Threats to Validity

7 Future Work

One of the possible future research objectives would be to test the framework on a real-world application, and to either enhance its use as an opportunistic goal-oriented application, or to extend its use to support opportunistic information discovery, information curation, or goal realization.

8 Conclusion

References

[1] Abrams, David, Ron Baecker, and Mark Chignell. "Information archiving with bookmarks: personal Web space construction and organization." *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co., 1998.

[2] Bates, Marcia J. "Toward an integrated model of information seeking and searching." The New Review of Information Behaviour Research 3 (2002): 1-15. APA

[3] Bates, Marcia J. "An exploratory paradigm for online information retrieval." *Intelligent Information Systems for the Information Society.* Amsterdam: North-Holland (1986): 91-99.

[4] Beagrie, Neil. "Digital curation for science, digital libraries, and individuals." *International Journal of Digital Curation* 1.1 (2008): 3-16.

[5] Choo, C. W., Detlor, B., and Tunbull, D. (2000). Information seeking on the web: An integrated model of browsing and searching. *FirstMonday*, 5(2). Available from http://firstmonday.org/issues/issue5_2/choo/index.html.

[6] Ellis, David. "A behavioural model for information retrieval system design." *Journal of information science* 15.4-5 (1989): 237-247.

[7] Ellis, David, Deborah Cox, and Katherine Hall. "A comparison of the information seeking patterns of researchers in the physical and social sciences." *Journal of documentation* 49.4 (1993): 356-369.

[8] Ellis, David, and Merete Haugan. "Modelling the information seeking patterns of engineers and research scientists in an industrial environment." *Journal of documentation* 53.4 (1997): 384-403.

[9] Foster, Allen, and Nigel Ford. "Serendipity and information seeking: an empirical study." *Journal of Documentation* 59.3 (2003): 321-340.

[10] Java, Akshay, et al. "Feeds That Matter: A Study of Bloglines Subscriptions." *ICWSM*. 2007.

[11] Kellar, Melanie, Carolyn Watters, and Michael Shepherd. "A Goal-based Classification of Web Information Tasks." *Proceedings of the American Society for Information Science and Technology* 43.1 (2006): 1-22.

[12] Kellar, Melanie, Carolyn Watters, and Michael Shepherd. "A field study characterizing Web-based information-seeking tasks." *Journal of the American Society for Information Science and Technology* 58.7 (2007): 999-1018.

[13] Levene, Mark. *An introduction to search engines and web navigation.* John Wiley & Sons, 2011.

[14] Lindley, Siân E., et al. "It's simply integral to what I do: enquiries into how the web is weaved into everyday life." *Proceedings of the 21st international conference on World Wide Web.* ACM, 2012.

[15] Locke, Edwin A., and Gary P. Latham. "Building a practically useful theory of goal setting and task motivation: A 35-year odyssey." *American psychologist* 57.9 (2002): 705.

[16] Mishne, Gilad, and Maarten De Rijke. "A study of blog search." *Advances in information retrieval.* Springer Berlin Heidelberg, 2006. 289-301.

[17] Morrison, Julie B., Peter Pirolli, and Stuart K. Card. "A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions." *CHI'01 extended abstracts on Human factors in computing systems.* ACM, 2001.

[18] Sellen, Abigail J., Rachel Murphy, and Kate L. Shaw. "How knowledge workers use the web." *Proceedings of the SIGCHI conference on Human factors in computing systems.* ACM, 2002.

[19] Whittaker, Steve. "Personal information management: from information consumption to curation." *Annual review of information science and technology* 45.1 (2011): 1-62.

[20] Yin, R. K. 2009. *Case study research*, 4th, Thousand Oaks, CA: Sage.