IDENTIFYING THE BOUNDS OF AN INTERNET RESOURCE

A Dissertation

by

FARYANEH POURSARDAR

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Frank Shipman
Committee Members,	Richard Furuta
	James Caverlee
	Ann McNamara
Head of Department,	Dilma Da Silva

December 2018

Major Subject: Computer Science

ABSTRACT

Systems for retrieving or archiving Internet resources often assume a URL acts as a delimiter for the resource. But there are many situations where Internet resources do not have a one-to-one mapping with URLs. For URLs that point to the first page of a document that has been broken up over multiple pages, users are likely to consider the whole article as the resource, even though it is spread across multiple URLs. Comments, tags, ratings, and advertising might or might not be perceived as part of the resource whether they are retrieved as part of the primary URL or accessed via a link.

Understanding what people perceive as part of a resource is necessary prior to developing algorithms to detect and make use of resource boundaries. A pilot study examined how content similarity, URL similarity, and the combination of the two matched human expectations. This pilot study showed that more nuanced techniques were needed that took into account the particular content and context of the resource and related content.

Based on the lessons from the pilot study, a study was performed focused on two research questions: (1) how particular relationships between the content of pages effect expectations and (2) how encountered implementations of saving and perceptions of content value relate to the notion of internet resource bounds. Results showed that human expectations are affected by expected relationships, such as two web pages showing parts of the same news article. They are also affected when two content elements are part of the same set of content, as is the case when two photos are presented as members of the same collection or presentation. Expectations were also affected by the role of the content – advertisements presented alongside articles or photos were less likely to be considered as part of a resource.

The exploration of web resource boundaries found that people's assessments of resource bounds rely on understanding relationships between content fragments on the same web page and between content fragments on different web pages. These results were in the context of personal archiving scenarios. Would institutional archives have different expectations? A follow-on study gathered perceptions in the context of institutional archiving questions to explore whether such perceptions change based on whether the archive is for personal use or is institutional in nature.

Results show that there are similar expectations for preserving continuations of the main content in personal and institutional archiving scenarios. Institutional archives are more likely to be expected to preserve the context of the main content, such as additional linked content, advertisements, and author information. This implies alternative resource bounds based on the type of content, relationships between content elements, and the type of archive in consideration.

Based on the predictive features that gathered, an automatic classification for determining if two pieces of content should be considered as part of the same resource was designed. This classifier is an example of taking into account the features identified as important in the studies of human perceptions when developing techniques that bound materials captured during the archiving of online resources.

DEDICATION

To my mother, to my father, to Ali

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Frank Shipman, and my committee members, Dr. Richard Furuta, Dr. James Caverlee, Dr. Ann McNamara for their guidance and support throughout the course of this research.

Many thanks to my mother, my father, and my sister who are always there for me, for their love, tremendous support and encouragement and to my husband, Ali, for his patience and love. I dedicate this work to my parents and my husband, to gratitude their love, patience and support during these years. Thanks to my precious little Sarah. I am so fortunate to have her, she is the hope in my life.

I am grateful to my friends and colleagues and the department faculty and staff for making my time at Texas A&M University a great experience. I also want to extend my gratitude to the members of the Center for Study of Digital Libraries; Anna Zacchi, Dohyoung Kim, Luis Meneses, Sampath Jayarathna, Suinn Park (Josh), Gabriel Dzodom, and Caio Duarte Monteiro. I will remember my lab-mates Elham Khabiri, Said Kashoob, Hamed Alhoori, Kyumin Lee, Amir Fayazi, Zhiyuan Cheng and Krishna Kamath for discussing several research topics, studying together and being my friend.

CONTRIBUTORS AND FUNDING SOURCES

This work was supervised by a dissertation committee consisting of Professor Frank Shipman [advisor], Professor Richard Furuta and Associate Professor James Caverlee of the Department of Computer Science and Engineering and Associate Professor Ann McNamara of the Department of Visualization.

All work for the dissertation was completed by the student, under the advisement of Professor Frank Shipman of the Department of Computer Science and Engineering.

There are no outside funding contributions to acknowledge related to the research and compilation of this document.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

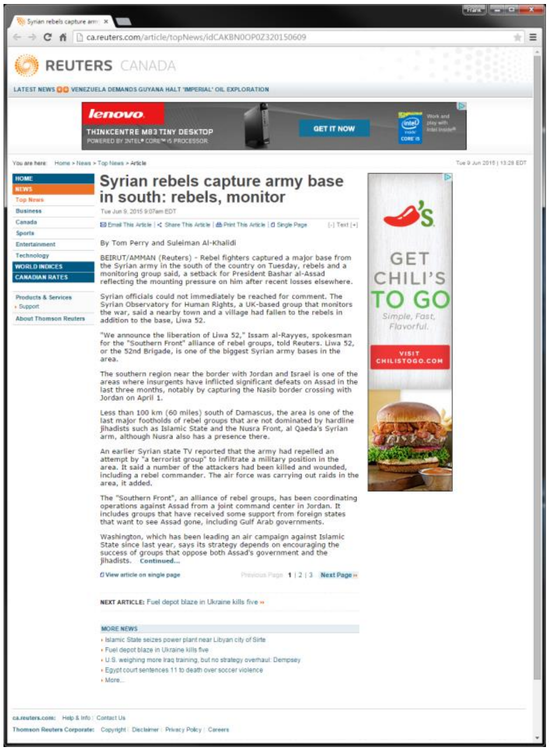
We live a world in which information is expected to be always at hand. Search engines and archiving tools mediate access to much of the content available on the Internet. Whether indexing the contents to enable search or determining what contents need to be saved for archiving, systems need an accurate model of what content is and is not part of a resource.

Systems for retrieving or archiving Internet resources often assume a URL acts as a delimiter for the resource. But there are many situations where Internet resources do not have a one-to-one mapping with URLs. For URLs that point to the first page of a document that has been broken up over multiple pages, users are likely to consider the whole article as the resource, even though it is spread across multiple URLs. Comments, tags, ratings, and advertising might or might not be perceived as part of the resource whether they are retrieved as part of the primary URL or accessed via a link. Similarly, whether content accessible via links, tabs, or other navigation available at the primary URL is perceived as part of the resource may depend on the design of the website.

Misidentification of resource boundaries results in false positives in search results when components of a web page unrelated to the main resource (e.g. advertising, off-topic comments) are indexed with the resource. But it also affects users through false negatives that occur when the contents of a resource are spread across multiple web pages. Figure **1** shows the first page of a news article spread across multiple URLs.

Without understanding what is part of the article, what is advertising, and what is site-oriented navigation, systems will incorrectly index or archive such resources.

Our work explores the question of how users perceive the bounds of Internet resources. We do this in the context of archiving, where it is more straightforward to ask users about what is and is not part of a resource. The next section further discusses how the bounds of resources can be challenging to define when archiving Internet resources. After this we describe our pilot and complementary studies and the findings in the following chapters. This leads to a discussion of implications and conclusion.


Syrian rebels capture arm
ca.reuters.com/article/topNews/idCAKBN0OP0Z320150609
REUTERS CANADA
LATEST NEWS VENEZUELA DEMANDS GUYANA HALT 'IMPERIAL' OIL EXPLORATION
lenovo
THINKCENTRE M83 TINY DESKTOP
POWERED BY INTEL® CORE™ i5 PROCESSOR
GET IT NOW
intel inside CORE i5
Work and play with Intel Inside®
You are here: Home > News > Top News > Article
Tue 9 Jun 2015 | 13:28 EDT
HOME
NEWS
Top News
Business
Canada
Sports
Entertainment
Technology
WORLD INDICES
CANADIAN RATES
Products & Services
Support
About Thomson Reuters
Email This Article | Share This Article | Print This Article | Single Page
[-] Text [+]
GET CHILI'S TO GO
Simple, Fast, Flavorful.
VISIT CHILISTOGO.COM
View article on single page
Previous Page 1 | 2 | 3 Next Page
ca.reuters.com: Help & Info | Contact Us
Thomson Reuters Corporate: Copyright | Disclaimer | Privacy Policy | Careers

Syrian rebels capture army base in south: rebels, monitor

Tue Jun 9, 2015 9:07am EDT

By Tom Perry and Suleiman Al-Khalidi

BEIRUT/AMMAN (Reuters) - Rebel fighters captured a major base from the Syrian army in the south of the country on Tuesday, rebels and a monitoring group said, a setback for President Bashar al-Assad reflecting the mounting pressure on him after recent losses elsewhere.

Syrian officials could not immediately be reached for comment. The Syrian Observatory for Human Rights, a UK-based group that monitors the war, said a nearby town and a village had fallen to the rebels in addition to the base, Liwa 52.

"We announce the liberation of Liwa 52," Issam al-Rayyes, spokesman for the "Southern Front" alliance of rebel groups, told Reuters. Liwa 52, or the 52nd Brigade, is one of the biggest Syrian army bases in the area.

The southern region near the border with Jordan and Israel is one of the areas where insurgents have inflicted significant defeats on Assad in the last three months, notably by capturing the Nasib border crossing with Jordan on April 1.

Less than 100 km (60 miles) south of Damascus, the area is one of the last major footholds of rebel groups that are not dominated by hardline jihadists such as Islamic State and the Nusra Front, al Qaeda's Syrian arm, although Nusra also has a presence there.

An earlier Syrian state TV reported that the army had repelled an attempt by "a terrorist group" to infiltrate a military position in the area. It said a number of the attackers had been killed and wounded, including a rebel commander. The air force was carrying out raids in the area, it added.

The "Southern Front", an alliance of rebel groups, has been coordinating operations against Assad from a joint command center in Jordan. It includes groups that have received some support from foreign states that want to see Assad gone, including Gulf Arab governments.

Washington, which has been leading an air campaign against Islamic State since last year, says its strategy depends on encouraging the success of groups that oppose both Assad's government and the jihadists. Continued...

NEXT ARTICLE: Fuel depot blaze in Ukraine kills five

MORE NEWS
- Islamic State seizes power plant near Libyan city of Sirte
- Fuel depot blaze in Ukraine kills five
- U.S. weighing more Iraq training, but no strategy overhaul: Dempsey
- Egypt court sentences 11 to death over soccer violence
- More...

er

Figure 1. First page of a multi-page article on the Web. Systems need to determine if the website navigation components, advertisements, related articles, and story pages 2 and 3 are part of the resource being indexed or archived.

1.1 What is an Internet Resource?

When we point to a resource, we expect that resource to be available as a whole. But what does it mean to be whole? When people point to the main page of a web site, do they expect the site as a whole to remain available? When pointing to an item on Amazon, should all the comments and ratings also be available? In contexts where the receiver has access to the Internet and where the resource is from a reliable provider, there is no issue. But when the reference is being used to either create a temporary offline version of the resource or to create a copy for long-term archiving [101], the question of which content is expected to be part of the resource becomes a central consideration. And the question of how much content is expected to be part of the resource becomes central to deciding how systems should behave.

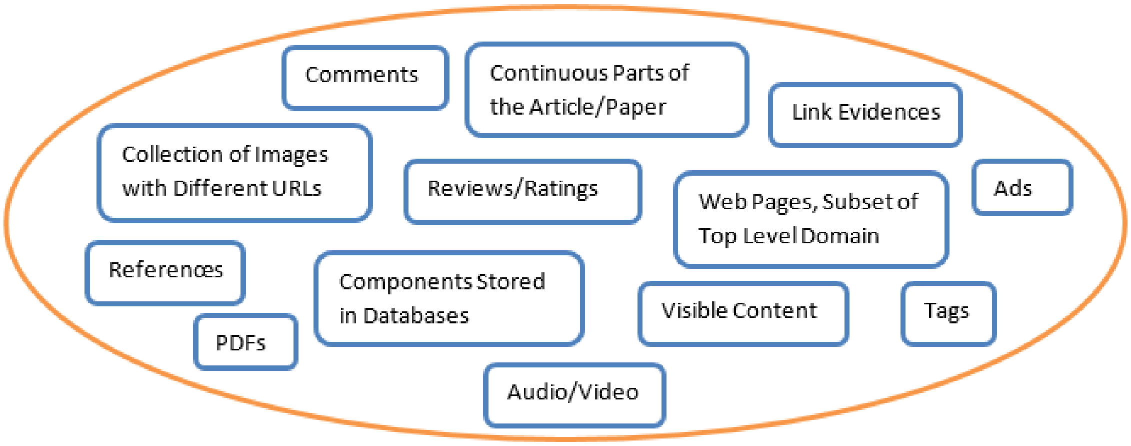

Comments
Continuous Parts of the Article/Paper
Link Evidences
Collection of Images with Different URLs
Reviews/Ratings
Web Pages, Subset of Top Level Domain
Ads
References
Components Stored in Databases
Visible Content
Tags
PDFs
Audio/Video

Figure 2.An Internet Resource can be a combination of various components

The answers to the above questions are complicated by the fact that the content visible on web pages is rarely fetched through a single http request (see Figure **2**). Web pages include frames and other methods for generating a page based on a variety of static and dynamic content. This content takes the form of links to external web pages, audio, video, advertisements based on user interest, comments, tags, likes, etc. When people refer to the web page as a resource, it is unclear how many of these components they are identifying as part of the resource. Also, it is dependent on the particular situation; most reuse contexts would not require the same advertising so be shown beside a news story with embedded images and video but there are contexts where this would matter. It might be the juxtaposition that was being preserved – such as recording the perceived irony of an advertisement for vacations in Florida appearing next to a story about a Florida hurricane.

There may also be expectations for related content. A resource whose identifying link is to the first page of an article is almost certainly expected to include the content for the remaining pages of the article. Similarly, the navigation bars on web pages can be designed to imply a single resource separated into components available through tabs even though they are retrieved through independent URLs. Some references to the resource would include the content on these tabs even though they are not visible when going to the initial URL. It is likely that strongly-related content embedded in a single web page or divided over a set of content across multiple URLs are likely to result in fairly consistent expectations by users. But expectations are likely to vary considerably for a web site with more loosely connected content. When someone points to the top page of such content, they may be indicating the value of the site as a whole, a subset of the site that is more conspicuously related to the top page, or just the initially visible content. Many classes of systems are impacted by this uncertainty about what constitutes the resource being referenced, resulting in a number of difficult issues for system designers/developers. Search engines, recommender systems, and archiving tools would all better meet their users' needs if what was being referenced by users was clear. The focus in this research work is on how this question is answered in archiving contexts. Archiving systems are designed to provide access to the content at the end of a URL when it is not otherwise available, either due to the user not having access to the Internet or due to the content not being available from the original source either temporarily or permanently. This dissertation studies how to design systems to identify these boundaries of a resource automatically. These boundaries are defined as the information components expected to be available by users. Thus, the first line of research to being able to algorithmically compute resource boundaries is to better understand what users' expectations are. Once we have a better understanding of user expectations, we develop and compare algorithms to heuristically determine resource boundaries.

1.2 Contributions

This dissertation explores the following research questions related to understanding and identifying the bounds of internet resources:

1. We investigate how characteristics of resources and relationships between related content affect user perceptions when identifying digital items which are conceptually bundled together in an Internet resource.

2. We compare how people’s responses differ to alternate statements when assessing resource bounds, including statements emphasizing the value of related material and statements directly addressing being part of the same resource.
3. We examine how the type of archive, personal or institutional, changes people’s assessments of the bounds of resources.
4. Finally, we provide an example of how the characteristics identified in the human studies can be used to create an automatic approach to assess resource bounds.

The work presented in this dissertation can inform the design of personal archiving tools and also the harvesting policies of institutions.

1.3 Overview of This Dissertation

The rest of this dissertation is organized as follows:

Chapter 2: This chapter presents related work, especially studies related to the challenges faced by individuals and organizations when archiving web content. Additionally, research exploring users’ perceptions of personal and archived digital content, practices surrounding how people preserve and manage their personal information, and a review of how institutional archives and personal archives have differed in terms of tools, practices, and perceptions.

Chapter 3: A pilot study of how systems might identify, or at least estimate, the boundaries of a resource automatically is presented. The pilot study asks people about the value of related content in the context of an archive with an emphasis on easily computable characteristics of pairs of web pages: having similar content and having similar URLs. The results help identify challenges to automatic boundary detection and lead to follow-on studies exploring more complex features of web page pairs.

Chapter 4: In this chapter, we continue our study exploring features potentially related to or part of Internet resources. The assessment is exploring user perceptions when more specific relations between the two pages are present or absent. The study focused on two research questions: (1) how particular relationships between the content of pages affect expectations and (2) how encountered implementations of saving and perceptions of content value relate to the notion of Internet resource bounds. The expressed user perceptions identified interaction between composite resource types and the relations between page pairs that influence their being considered as part of the same resource. Based on the results of the preliminary study presented in Chapter 3, three different questions were asked from the participants in the user study to understand their perception on the boundaries of Internet resources. The variations in answers for the same page pairs to the different questions provides an understanding of the relative frequency of perceiving value for a related page, perceiving that page to be part of the same resource, and expecting that it will be archived when the first page is archived.

Chapter 5: In this chapter, we conduct a study to find the answer to the question of how expectations change depending on the type of archive in question. In particular, we explore how perceptions change when changing the archive from a personal context to an organizational archive. The results of the two studies on personal and institutional archives show that there is a greater expectation for content that is not clearly related to the primary content to be archived in an organizational context.

Chapter 6: This chapter describes the design of classifiers based on the results of the studies presented in Chapters 3-5. This provides a model of how future system building can make use of human perceptions when considering the design of algorithms that determine the boundaries of an archive. Two classifications are identified and explored: “*Single_Resource*” and “*Resource_Type*”. These determine whether two pages are part of a single resource and what is the overall type of the resource. The results are compared using different classification techniques using the features developed based on the studies and results using these features is compared to using only the content similarity and URL similarity originally hypothesized as being valuable to this process.

Chapter 7: We conclude with a summary of the contributions of this work and present a discussion of future research possibilities.

2. LITERATURE REVIEW

There has been a growing interest in how people value or keep digital things. Information on the Web becomes a thing to be addressed, linked to, organized, and placed into larger contexts. Fetterly et al. [33] estimate that about two percent of the Web disappears from its current location every week. To have access to the content of some of these missing pages, entities such as the Internet Archive (IA) [101] select content to save in case of its removal. Additionally, as part of the process of indexing web content, search engines also create temporary cached versions of web resources.

To have easy access to our digital belongings, there are different approaches. One line of research views the problem as one of identifying moved or changed resources and refinding or assessing the types of change to these resources. This view results in systems-oriented solutions. For example, Phelps and Wilensky [89] and Dalel et al. [24] investigate the effectiveness of saving identifying search terms and phrases respectively for relocating moved resources or potentially finding appropriate replacements. Similarly, systems have been developed that try to regenerate missing resources. For example, Warrick [79] attempts to restore lost web resources through reconstruction [78, 80].

A second line of research views the problem as one of personal information management or personal archiving [47]. In this view, users directly or indirectly identify content they wish to have later. While such an approach requires enabling systems, there is also a need to understand user practices and desires. As such there is a body of literature examining how people think of their digital contents and how they address needs and problems with these contents as they arise.

2.1 A System View: Challenges for Web Archives

Web archives attempt to preserve the fast changing Web, yet they will always be incomplete. Due to restrictions in crawling depth, crawling frequency, and restrictive selection policies [45], large parts of the Web are unarchived and lost. If archive boundaries are drawn without an understanding of the content, this loss will include scientific, cultural and other valuable content. Even so, the captured pages might be deprived of certain elements. Bar–Yosseff et al. carried out experiments to measure the decay of the Web [6]. SalahEldeen determined that nearly 11% of shared resources will be lost one year after being published and that this decay will continue at a 0.02% rate per day [94]. Crawlers often fail to capture embedded elements; JavaScript and Flash are among the examples [28, 44, 77]. Brunelle et al. [17] have showed that the Internet Archive is missing an increasing number of important embedded resources over the years. The work presented here aims to help determine what is and is not considered part of a resource.

Some previous work on finding missing resources is based around the premise that documents and information are not lost but simply misplaced [4] as a consequence of the lack of integrity in the Web [3, 26]. Other studies have also focused on finding the longevity of documents in the Web [52] and in distributed collections [61, 98].

2.1.1 Inaccessible Web Resources

Inaccessible Web pages and “page not found” are common issues that arise during Web browsing. However the information on the Web is often not lost completely, it can be moved from one URL to another [82] for numerous reasons (e.g. Web site moved or reorganized.) Usually the whole or just part of the content is available [58]; Even when the original content provider removes content from the Web, some or all of a resource may still be available. In [58] [59] [60] authors propose four retrieval methods for discovering missing Web pages using: 1. lexical signatures (LS), 2. Web page titles, 3. social bookmarking tags, and 4. link neighborhood lexical signatures (LNLS). A lexical signature consists of light weight metadata representing the content of a document. Using the title of a Web page to rediscover the missing page may work as titles are descriptive and rarely change over time. Tags are terms suggested by users as describing the content of the page, such as tags collected on delicious.com [57, 59]. LNLS is a lexical signature generated from the pages that link to the missing page rather than or in addition to the page itself [60].

Another line of work focuses on methods for detecting and categorizing changes in Web documents within a collection. The aim of this work is to offer strategies that can be implemented to effectively detect and alleviate the consequences of the various types of change [83]. In addition to detecting change, these systems may also address the issue of missing Web pages and curating missing resources [35]. Furuta et al. developed Waldens Paths, a tool which allows users to construct trails using Web pages which are usually authored by others [36]. This path can be seen as a meta-document that organizes and adds contextual information to those pages. Thus, as part of aiding path management by path authors, the research also explored discovering relevant and significant changes to websites. In order to infer change relevance, Francisco-Revilla et al. [34] developed techniques to identify and visualize the nature of the Web page change in four categories: content or semantic, presentation, structural, and behavioral. These techniques were based on monitoring the document signatures of paragraphs, page headings, links out of the page, and keywords.

In another study, Meneses et al. [82, 83] explored one specific type of page change that is particularly problematic for archives. This work showed that soft 404s, error pages that are not reported as such by Web servers, can also be identified with text classifiers based on the characteristics of previously identified soft 404 pages. The authors were able to isolate lexical signatures of such pages, which contributed to predicting soft 404s with a precision of 99% and a recall of 92%. The work has focused on restoring the semantic integrity of incomplete document collections [81].

2.1.2 Other Problems and Complications

Constant modifications of data, changes to Web services, and intermittent technical problems make preserving Internet resources challenging. How does the archiving site need to prepare the website for preservation? How can metadata be derived for Web resources? These are questions that have to be answered before being able to archive a resource. The site preparation process is challenging, and an archive needs to process each resource with metadata extraction utilities to record content and technical information.

Some preprocessing can occur at the originating server. Modules, like CRATE [96], enable the Web server to process complex object metadata formats so that the Web server responds to the archiving repository crawler by sending both the resource and the just-in-time generated metadata [97]. For instance, to better preserve an image, the archive needs supporting data such as subject matter or what the picture is about, its creator and origination hardware (camera type) and software. CRATE and similar modules can provide such metadata which is not provided normally due to limitations of MIME typing as currently implemented by Web servers.

Smith and Nelson [96] present several tools that can be helpful for processing resources in preparation for digital preservation. For example, the Global Digital Format Registry (GDFR) [2] and Pronom's [25] DROID tool provide a deeper reflection of the resource's format. JHOVE [46] can identify, validate and characterize a number of file types including images, text, and PDF documents.

In a research, Decman [29] explains how constant changing of the data causes constant loss of a huge amount of information and the loss of scientific, cultural and other heritage. And all of these happen because of technical problems of long-term preservation of Web pages [29]. In recent years, the Web has become a nexus for sharing, publishing, and storing personal content. In other words, the Web has increasingly become a place where much of our personal content is archived. This raises the question of what kinds of tools do people need to manage, maintain, and keep track of the content that matters to them on the Web? One part of the current research aims to inform the design of personal archiving tools to respond to the user needs.

2.2 A Human View: User Perceptions of Virtual Possession

Web sites and services are not static: a service can be shut down or an account can became inactive. Indeed, loss on the Web has been attributed to many sources other than technology failure [47] [75]. Recent research has indicated that the lack of certainty as to where content that is hosted online really is has consequences for one's sense of control and ownership over it [86]. As a result, people may download content onto laptops and other local devices to create a stronger sense of ownership [86]. However, these approaches mean that potentially interesting digital attributes and social metadata, such as comments, tags, and likes, may be either lost or fixed in time. Giving users the ability to feel in control of online content [41], while at the same time being able to retain some of its valuable on-line attributes, is related to the problem of recognizing relationships between primary resource content and its social annotation through community activity.

Marshall's study [73] introduces four challenges for personal digital archiving presented as:

1) Accumulation: it's hard to give value to digital belongings and separate them based on it.

2) Distribution: a person's digital belongings are distributed among different stores for a variety of reasons. Most common reasons that people replicate files can be: a) to back up important files against immediate loss; b) to share them with other people; or c) to use online files locally.

3) Digital management: people are unwilling to spend very much energy on curation, while willing to take advantage of relying on the existing facilities to keep their digital assets safe.

4) Long-term access: sometimes we don't remember what we have or where the digital assets are or even changing the technology prevents us from having access to our belongings [72].

To overcome these challenges we need to answer questions like: What should we keep? Where should we put it for later use? How should we maintain it? And what do we expect to have access to after saving them?

It is easy to accumulate digital belongings. But is it worthy to gather everything? You need to spend a lot of time to review and harvest useful and important materials. The notion of drawing all digital assets together and having a centralized personal archive store is not practical today. In a study, informants exhibit that they have a variety of options to save their belongings such as on a local or network store, removable drive; or using network storage, the "storage in the cloud" solution [66].

The standard view with ‘the Cloud’ is that people will be able to keep their digital assets more securely and more cheaply. By moving away from local storage, users can be sure that when their devices crash or get stolen their data will be safe. Odom et al. research [86] focuses on the concerns of the users about where and how to keep their digital material. Several factors shape participants’ motivations to put their personal digital stuff on Internet.

What interviews reveal [86] is that keeping things online in some sense hands that accountability over to some unknown, unseen entity and further that people may have very little faith to it. It is no wonder that as users of contemporary technology increasingly engage with their digital stuff, seeking to place it in secure storage, sharing it with others, and sometimes wanting to know ‘who has it’ or ‘where it has gone’, that they end up personalizing their own versions and preserving them in a secure place for later access.

There is a line of research that helps us to better understand people’s archival behavior and expectation also illustrates how they manage these belongings. Considering how people preserve and manage their virtual possessions can give us a point of view in their archiving behavior and is a helpful hint on identifying the bound of the Internet resources. With this assumption we studied existing research in the area of Personal Information Management (PIM).

2.2.1 Personal Information Preservation and Management

Jones [32] defines PIM as both the practice and the study of the “activities a person performs in order to acquire or create, store, organize, maintain, retrieve, use and distribute the information needed to complete tasks and fulfill various roles and responsibilities” (p.453).

Henderson has looked at how people organize their desktops [42] and Jones et al. have reported on the extensive use of desktop folders [51]. Other researchers have studied in detail the use, archiving and/or storage of emails [10-12, 31, 104], of documents [15], of time management tools [13], of To Do lists [9] and personal information management software [9, 14, 56, 87].

PIM studies have found that maintenance and organization of information is less of a priority to individuals than time-sensitive and context-driven activities such as finding and keeping [8, 16]. Maintenance activities include storing, organizing, deleting, and reorganizing information. In the PIM literature, there is limited research on the behaviors related to digital preservation. The studies by Marshall, et al. [70, 75] and Marshall [70], [73], [72] identified issues faced by technically knowledgeable computer users when dealing with long term access to email and their own digital files.

The study of self-archiving of electronic personal records by scholars, artists, academics, and politicians has been the focus of the archival field [54], [55], as is the ingestion of personal electronic records into existing institutional archives [22], [27].

Marshall et al. [74, 76] have conducted a field study to understand how individual consumers acquire, keep, and access their digital belongings with the focus on determining of what they had kept, which of these belongings are more important to them over the long term, and what difficulties they had in maintaining them. The study reveals that the backup and file replications are the major common belief among individuals for long term archiving, but these are several contradictions: for example even though all informants agreed that replication is a valuable safety net in principle they barely replicate their files on CDs, removable devices. It is notable that most of them use ad-hoc products for their simplicity of use (e.g. archiving as an email attachment).

Advances in storage capacity and reduction in cost inspire the consumers to keep all the data but the desire of having control on storage volumes cause them to delete some data which seems unimportant to the consumers.

In another study, authors [74] addressed central challenges for personal archiving and categorized them in four groups: 1) digital materials accumulated in a different problematic way than physical materials; 2) digital materials are fundamentally distributed; 3) standard curation problems such as managing files in aggregate, creating appropriate metadata are magnified in the consumer setting; 4) facilities for long-term access are not supported log in the current desktop metaphor.

There are different suggestions for keeping and managing personal information. Odom et al. [86] propose an archive collection to be constructed by unifying content in a virtual way from many distributed online sources, so that it can be viewed and managed as a whole. Alternatively, Marshall [73] suggests such an archive collection be created by federating metadata records in a centralized store until the actual online venues disappear, so that the content can be automatically saved at that time. Both solutions raise a deeper question of federation: does it make sense to bring distributed online personal resources back together as an archive? Is it valuable to be able to view and manage distributed online resources together? In a different study, Lindley et al. [66] have conducted a study to understand whether a sense of ownership and control can be reinforced by unifying online content as a virtual, single store or not. The analysis shows that notion of drawing it together, in a secure and centralized archive is impossible.

Several studies have focused on the effectiveness and potential of tools to aid in organizing and re-finding of personal information. The five major areas under study are: Web site management tools [1], [38], [30], [16]; files and folders [18], [8], [7], [51]; email [104], [5], [31], [71], [102]; photographs [93], [23], [103], [88]; and cross-tool studies [14], [100], [53], [32]. The collective goal of the aforementioned research is to understand how these tools are used by individuals to manage personal information and to make recommendations for system and/or tool improvements based on those observations.

Jones et al. [48-50] discuss strategies people use to manage information for reuse. They address the problem of "Keeping Found Things Found" in other words once information found on World Wide Web, how are things organized for re-access and re-use later on? What can be done to avoid the need to repeat the process by which the information was found in the first place? They conducted an observational study of methods used in a workplace setting by users to manage Web information for re-use. The results of the study [50] showed that people observed in the study used a diversity of methods and associated tools. The outcomes also illustrated that several functions appear to influence the choice of method, included are the methods help Web information to be accessible from several places, methods help to share the information with others, or ones that remind the user of a Web page's relevance later on.

Furthermore, research has suggested that users have a weakened sense of possession over content that is stored in the Cloud [86] and want to personalized them by preserving them for later use. Harper et al. [41] point to social media as things that one might wish to download or otherwise act upon, but that do not support the simple range of actions normally associated with files. One cannot, for example, simply save a status update as a standalone object, or copy a photo that integrates the social metadata that is associated with it. Harper et al. suggest new actions are needed, which better enable users to act upon, and thus feel in control of, their online content. They believe such actions are essential if users are to have a greater sense of control, and ability to manage, digital content in a socially-networked world.

2.2.2 Perspectives on Institutional Archives

On the other hand, it is not only people who like to preserve online information; institutions like to archive the Internet resources as well. Institutional repositories provide managed access to the digital resources which produced and self-archived by the members of an institution [43]. In [68] Lynch point of view an organizational stewardship of digital resources consist of long term preservation, organization and access or distribution. The most important benefit of an institutional commitment is that a managed environment provides a greater degree of assurance of continued access than personal Web sites [99]. In [67] authors offer a three-level activity for preservation: a) Curation: the activity of managing and promoting the use of data from its point of creation, to ensure it is available for discovery and re-use. b) Archiving: a curation activity which ensures data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity. c) Preservation: an activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology. Cornel [20] believes “Access is still not the primary purpose of a preservation system”. Access is the purpose of an institutional repository; while preservation has a role in assuring long-term accessibility of the contents of an institutional repository.

For some institutes archiving and preservation of resources can relate to a historical perspective, like in museums, especially museums of history. Many studies are available on the long-term storage of digital information in such institutions [39, 63, 84, 105]. The work has been driven largely by libraries, museums, and governmental institutions. A benefit of these systems is accessibility; the data is searchable and can be quickly retrieved.

Institutional archiving requires an ongoing commitment to manage records to keep them intact and ahead of any type of time-dependent changes. Digital archiving systems promise to advance the ability of museums to preserve and utilize information about historical and cultural materials. Museums have a large number of materials that require preservation from future degradations for a long period as long as possible. On the other hand, it is common, and important to use these materials for a variety of purposes [64, 84], i.e., exhibitions, investigations, researches, education, and so on.

In museums of history, digital archives enable more flexible exhibitions that satisfy visitors' needs, deeper investigation and research, and a more convenient means of managing materials in comparison with conventional archiving systems. They also promise to increase accessibility to materials at lower cost [84].

Institutional archive increase accessibility to materials at lower cost and keeps the material for access by the next generation; although the purpose of it may be different from a personal archive [66]. Personal archive provide insight into what content is valued by users, and users directly or indirectly identify content they wish to have later as their own digital belongings [49, 75].

To sum up, the similarity and the difference between personal and institutional archive can be expressed in the following sentences. A personal digital archive usually is thematic [37] or subject-oriented, in contrast institutional archives capture the research and other intellectual, cultural or historical property generated by an institution's population active in many fields. Institutional digital archives often serve the purposes of preserving materials pertaining to the institution's history and to the activities and achievements of the institution [21]. A digital institutional archive can be any collection of digital material hosted, owned or published by the institute. The reason and answer to the question "why archive" is common in both personal and institutional settings. They both want to preserve important resources [54] and hope that they can retrieve the material on demand.

3. PRELIMINARY STUDY: EXPLORING THE EFFECTS OF CONTENT AND URL SIMILARITY[*]

The original motivation for this dissertation came from a practical problem encountered in the Walden's Paths project. That project developed a tool that enables users to preserve and provide access to the Internet resources included in paths, called PathCompiler [36]. PathCompiler was originally developed to let a variety of users "freeze" a set of Web pages for later use by others (see Figure **3**). This was a simple first approach for coping with change to resources [34, 35]. When users point to a resource by its URL, PathCompiler saved all necessary content of a Web page on the local machine along with the context necessary in order to show the Web pages off-line. Since the modern Internet is not just static pages, but is full of multimedia materials, pictures and scripts, PathCompiler traversed the links within a page to retrieve embedded content.

Upon initial use, it became clear that when users pointed to a resource via a URL they might mean only that page but often times they mean to provide access to a set of pages. Because of the challenge of identifying the bounds of a resource, a feature was added to PathCompiler to capture content at a number of links away from the original URL. But this approach is not efficient – the amount of network traffic and storage required is a function of the number of links on a page raised to the power of the chosen

[*] Reprinted with permission from "On Identifying the Bounds of an Internet Resource" by Poursardar, F. and Shipman, F., 2016. ACM Conference on Human Information Interaction and Retrieval, pp. 305-308, Copyright [2016] by ACM.

maximum link distance. If there are on average 30 links on a page and the distance selected is 2, there are on the order of 900 additional resources being archived for each original resource.

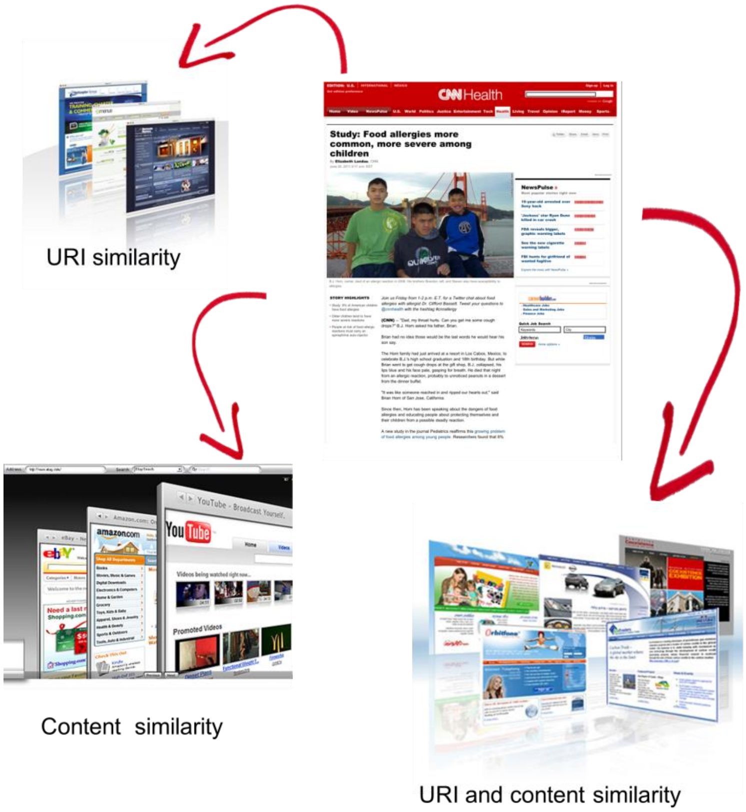


Figure 3.PathCompiler, a tool that enables users to archive Internet resources.

Thus, PathCompiler needed techniques to determine what the likely bounds for a resource are. The following heuristics were considered for archiving resources linked to from a Web page due to their ease of implementation:

1) Save linked materials that have content similar to the originally referenced resource,

2) Save links that have URLs similar to the referenced resource, and

3) Save links which have both similar content and similar URLs.

3.1 Study Approach and Method

To compare these alternatives, a study was performed to identify patterns in user expectations and desires when archiving resources. We asked 110 participants to indicate the value of archiving a second page when archiving an initial page. The pages were selected with an eye towards features likely to be part of an automatic approach to identifying resource bounds.

Table 1. Original primary Web page resources

Subject	Base URL
Technology	http://www.economist.com/news/science-and-technology/21573089-ambitious-project-map-brain-works-possibly-too-ambitious-hard
Daily news	http://shine.yahoo.com/love-sex/teenage-sweethearts-prove-it-s-never-too-late-as-they-and-reunite-and-marry-in-their-70s-163409359.htm
Health	http://www.naturalnews.com/033414_cancer_cures_documentary.html
Business	http://www.economist.com/news/briefing//21574489-britain-has-many-options-providing-extra-airport-capacity-its-capital-going-need

3.1.1 Corpus Development

For the purpose of this pilot study, we developed a corpus of four groups of Web resources in topic areas shown in Table 1. All selected Internet resources were in English. Using PathCompiler we crawled each original source page (the Base URL in Table 1) to extract all the links in the page. The contents of these linked pages constituted the corpus of potentially co-archived pages for assessment for that resource. Each of these pages was then categorized as being either similar in content or not and similar in URL or not. The cosine similarity of the term vectors for the original resource and the linked page (with a threshold of 0.7) was used to categorize them as having similar content or not. Resources from the same web site (same root URL) were considered to have similar URLs. Better techniques could be used for both classifications but initial results with these simple techniques could indicate where such effort should be spent.

3.1.2 Participant Selection

Our one hundred and ten (110) participants were identified among friends, colleagues, neighbors, family members, friends of friends, etc. Participants ranged in age from 26 to 70 and had a wide range of educational levels.

3.1.3 Participant Tasks

Each of the participants assessed 16 page pairs. For each pair, participants were simultaneously shown the primary resource and the potentially valuable content. They were asked to rate on a 5-point Likert scale – extremely useful, very useful, somehow useful, slightly useful, not at all useful - the value in archiving the second page. In total, 1760 web page pairs were assessed which is the sum of all ratings, shown in Figure 4 through Figure 7. The details about the assessments are shown in Table 2 and Table 3. Participants were not shown anything other than pair pages, no URLs and no data about how the pages were related.

3.2 Findings

The results show that people have different expectations based on what the original content is and how it is presented. Figure 4 through Figure 8 show the results for the four data sets. As is apparent from the highly varied distributions, each topic had a unique outcome. Where the health topic (Figure 4) had a relatively flat distribution of assessments across the five ratings, the technology topic distribution was heavily skewed to the negative (very few pages were viewed as part of the resource) and the daily news topic was classically bimodal with nearly all pages rated at the extremes.

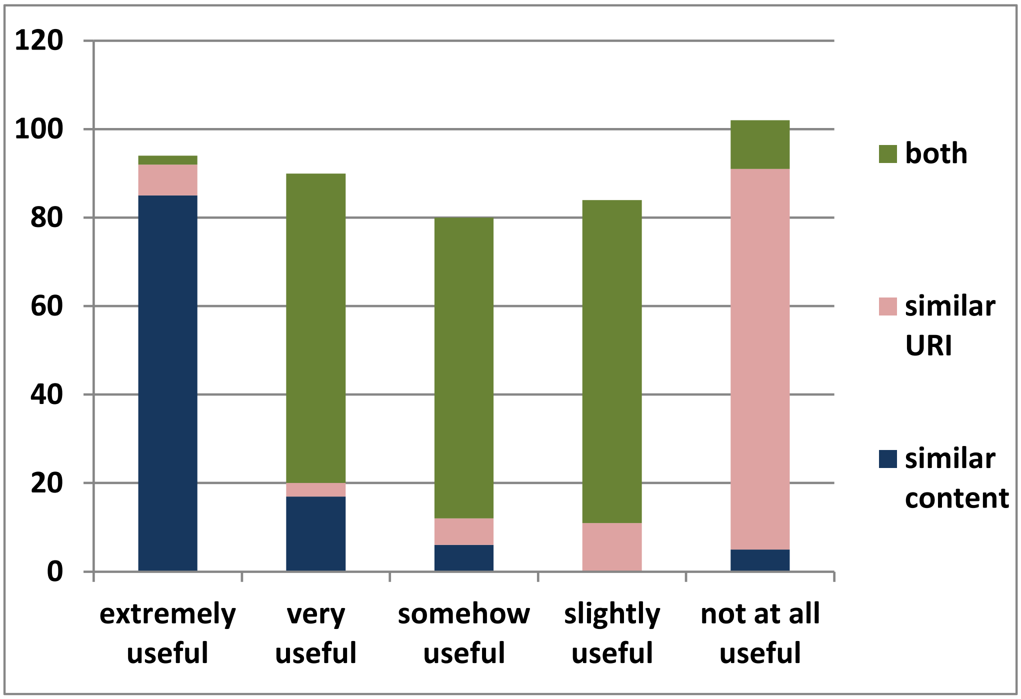


Figure 4. Similar content is more important than similar URL for health pages.

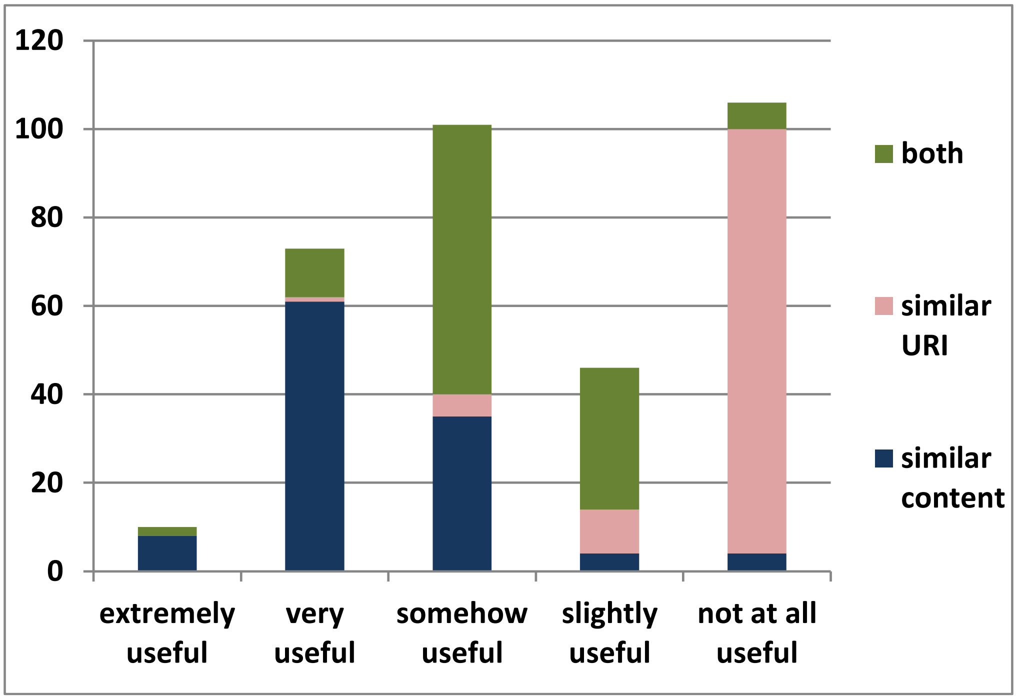


Figure 5.Similar content is more important than similar URL for business pages.

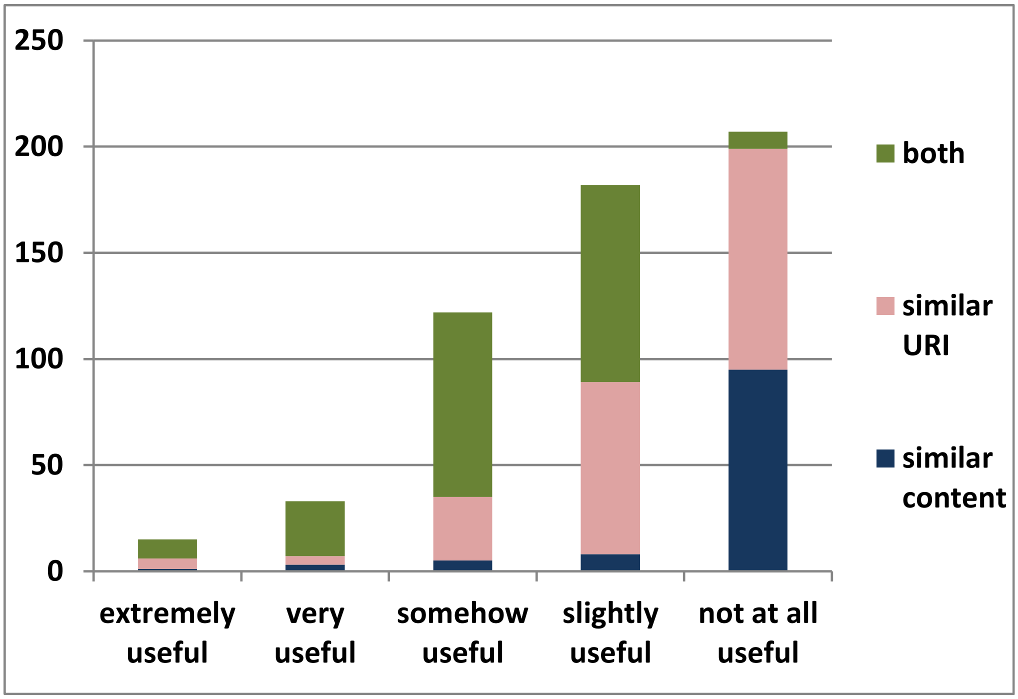


Figure 6.Not much was considered similar for technology pages, but most of what was had both similar URLs and similar content.

The data in Figure 4 through Figure 7 also shows which features correlated with high ratings vary across the four topic areas. While the highest rated content for the daily news group was overwhelmingly similar content from other servers, the ratings of the health and business topics were comparable among the pages with similar content on other servers and similar content from the same server (although in both cases the content from the same server was rated as slightly less valuable). For the technology topic, there was a strong preference for materials that were from the same server.

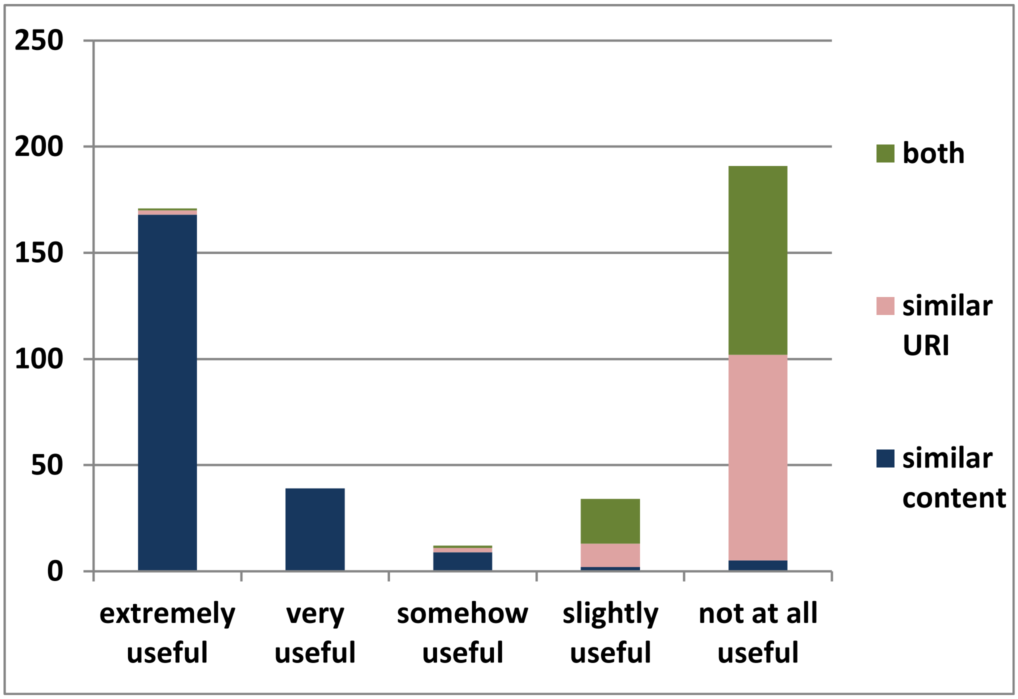
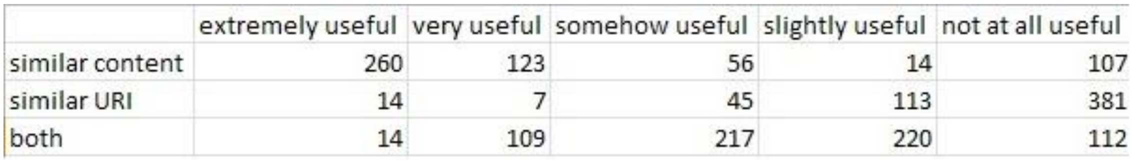


Figure 7.Similar content is more important than similar URL for news pages.

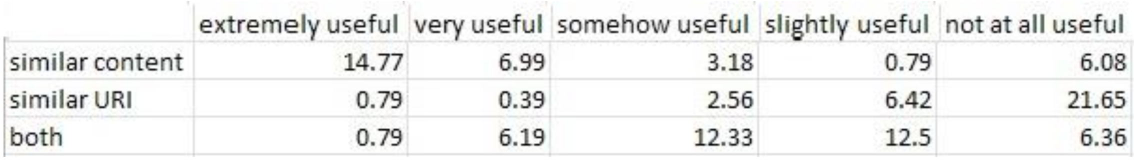
The following table (Table 2) shows the participants assessments (raw results) for all seed groups, based on the five Likert scales. For example, from the first row, 260 participants rated pages “extremely useful” when there was textual similarity between the resource and the potentially related material. A surprise result is the more negative reaction to similar content from the same site than reactions to similar content from other sites (i.e. similar content but with a dissimilar URL.) Numbers in Table 3 and Figure 8 show the finding the percentage form.

Table 2. Participants' assessments for all seed groups, based on the five Likert scale.



	extremely useful	very useful	somehow useful	slightly useful	not at all useful
similar content	260	123	56	14	107
similar URI	14	7	45	113	381
both	14	109	217	220	112

Table 3. Participants' assessment in the study in percentage.



	extremely useful	very useful	somehow useful	slightly useful	not at all useful
similar content	14.77	6.99	3.18	0.79	6.08
similar URI	0.79	0.39	2.56	6.42	21.65
both	0.79	6.19	12.33	12.5	6.36

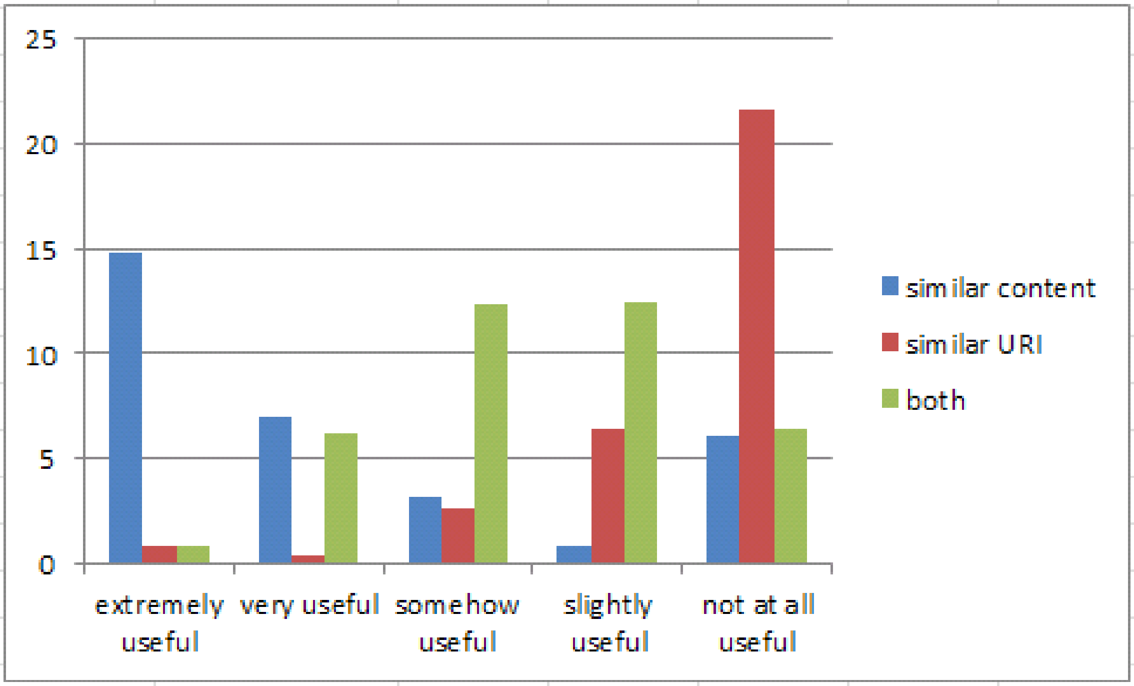


Figure 8 Participants' assessments for all seed groups, percentage form.

Overall, the results show that, while there was a relatively strong overall preference in preserving similar content, whether the URL of that content mattered varied across the four data sets. The primary lesson for those developing systems that preserve web-based resources from this study is that there is no simple answer to what is related to a resource.

3.3 Discussion and Conclusions

What might be going on? One answer is that participants intermixed information value assessments with their ratings of expectations of having the content available – and this was exacerbated by the imprecise wording of the Likert-scale statement in this preliminary study. The second answer is that the features that make a difference in whether people expect to have access to content are more nuanced than simply having similar content or a similar URL.

The results of this study show that content similarity is likely a viable feature for systems when deciding the bounds of a resource. The results also show that further study is needed to help design techniques to automatically identify the bounds of a resource.

While our focus has been on identifying the bounds of a resource for the purposes of archiving, the results of such investigations have broader implications. Search engines and recommender systems also benefit from more accurate assessments regarding Internet resource boundaries due to potential improvements to the content used when developing the indexes and content models used for retrieval. We hope this pilot study leads to greater interest in the dual challenges of determining what users perceive as the bounds of resources and techniques for systems to determine such bounds.

3.4 Summary

The preliminary data collection effort provided insight into issues in the design of data collection activities and infrastructure. Particular challenges include ensuring study participants are clear about the assessment they are making – it was noted that there was a confounding of whether participants were rating the independent value of the second page, the value of the second page relative to the first, or the expectation that the second page would be saved when the first page is identified as a resource to be archived/saved. As a result, the data collected from this process cannot be used for ground truth data about user expectations.

Despite these limitations, the study results indicate that the features important to this decision likely vary considerably from resource to resource. The study assessed two features of the pairs of pages: their textual similarity and the URL similarity with a small number of page pairs. The lack of a consistent connection between these two features and a positive assessment of the second page implies that additional features are needed to better understand and match user expectations [91]. The following chapter describes a more comprehensive study that considers different features of web page pairs.

4. MAIN STUDY: EFFECTS OF RELATIONSHIPS AND MEDIA TYPES IN A PERSONAL ARCHIVING SCENARIO[*]

Based on the lessons from the preliminary study [91], a larger data collection effort was undertaken to explore the effect of more specific features of web pages and relationships between them. The prior study indicated that URL and content similarity were not effective in predicting user perceptions about the bounds of resources. Our interpretation of this result was that more specific relationships between pages were dominating user responses. As a simple example, pages that showed parts of the same article would be considered part of the same resource but articles on the same topic from the same source (so similar in both content and URL) would not.

A second, unexpected, result of the preliminary study was the insight that people's understanding of the bounds of a resource may be affected by their prior experience with the "save" feature on software, and by their assessments of the value of content.

As a result, this larger study focuses on two research questions: (1) how particular relationships between the content of pages effect expectations and (2) how encountered implementations of saving and perceptions of content value relate to the notion of Internet resource bounds [92].

[*] Reprinted with permission from "What Is Part of That Resource? User Expectations for Personal Archiving" by Poursardar, F. and Shipman, 2017. ACM/IEEE Joint Conference on Digital Libraries (JCDL), pp. 229-232, Copyright [2017] by ACM/IEEE.

4.1 Study Approach and Method

To provide insight into these two questions, the study asked participants to look at page pairs where there was some known relationship between the two pages (e.g. there was a link from the first to the second page) and agree or disagree with the following three statements:

S1: If I save main page, I expect to save this too.

S2: If I have the first page, I would like to have access to this too.

S3: This is part of the resource so I would expect to have access if I save the main page.

The three statements are interrelated. S1 emphasizes the concept of saving that people encounter in web browsers. S2 emphasizes the likely value of having access to the second page while S3 asks about whether the second page is part of the same resource.

To explore the effects of different types of relationships between pages, four categories of content were explored: multi-page stories, image collections, reviews, and traditional web pages. To reduce the likelihood that results would be due to the particular content or idiosyncratic nature of the examples, 5-7 groups of web pages were captured for each of these four types of content.

Each group enabled the exploration of alternative relationships between pages of contents. Table 4 to Table 8 show the types of relationships included in the groups of pages and the number of instances of the relation in the group. There were a total of 122 page pairs presented to participants.

Table 4. Resource Types Considered in the Study

Type of Resource	*Number of Groups*
Multi-page story resource	5
Sequence of images resource	7
Reviews and ratings resource	5
Short single pages resource	7

Table 5. Page Relations for Multi-Page Story Resources

Considered Relations	*Number of Pages that Have the Relation*
Continuous pages of multi-page story	12
Links in the first page (main page)	5
Advertisements in the main page	3

Table 6. Page Relations for Set of Images Resources

Considered Relations	*Number of Pages that Have the Relation*
Images in the image set	31
Links in the main page/image	10
Author/photographer page	5

Table 7. Page Relations for Product Review Resources

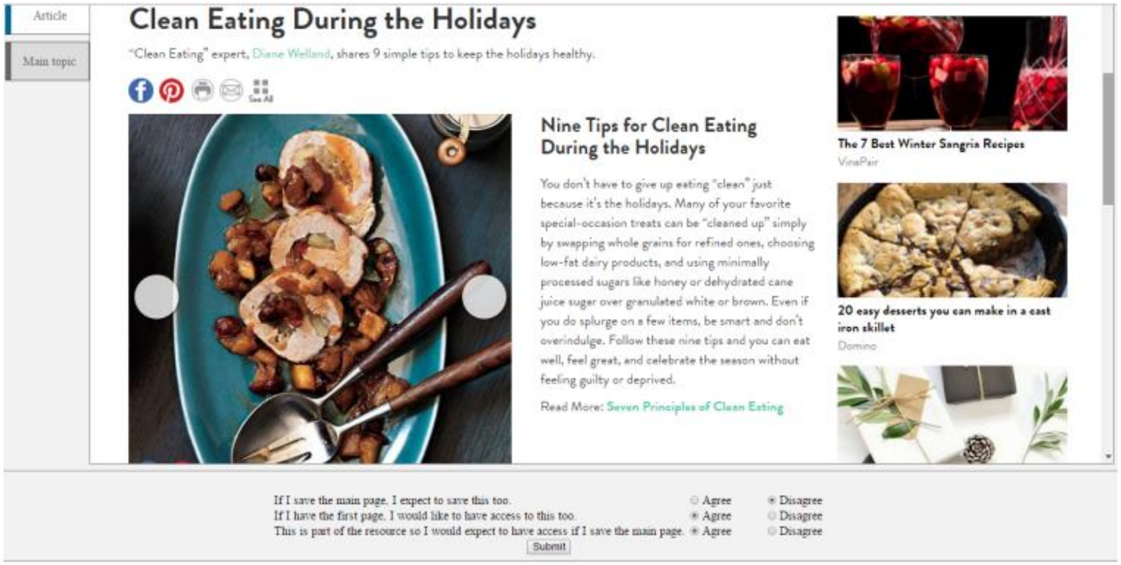
Considered Relations	*Number of Pages that Have the Relation*
related items/product/links pages	9
product Q&As pages	3
wish list/provider	7
reviews/ratings/comments pages	5

Table 8. Page Relations for Traditional Web Page Resources

Considered Relations	*Number of Pages that Have the Relation*
Links in the main page	14
Author page	6
Advertisements in the main page	12

To ensure consistency of content across the study, static versions of all components of the web pages in question were downloaded and cached. Each participant is asked to rate several pair pages where they are to assume they have asked to have access to the resource identified by the first page, then they were asked whether they expect to have access to the second page or not.

We developed a web interface for the study and used Amazon Mechanical Turk to recruit 2071 participants. Once a Turker accepted the task they were redirected to the study website. After examining the pair of pages and agreeing or disagreeing with the three statements for ~12 page pairs as shown in Figure 9, they were given a code to submit within Mechanical Turk for compensation. Not all users rated/assessed all the page pairs they were assigned but each page pair was assessed by 200-250 participants.



Clean Eating During the Holidays

"Clean Eating" expert, Diane Welland, shares 9 simple tips to keep the holidays healthy.

Nine Tips for Clean Eating During the Holidays

You don't have to give up eating "clean" just because it's the holidays. Many of your favorite special-occasion treats can be "cleaned up" simply by swapping whole grains for refined ones, choosing low-fat dairy products, and using minimally processed sugars like honey or dehydrated cane juice sugar over granulated white or brown. Even if you do splurge on a few items, be smart and don't overindulge. Follow these nine tips and you can eat well, feel great, and celebrate the season without feeling guilty or deprived.

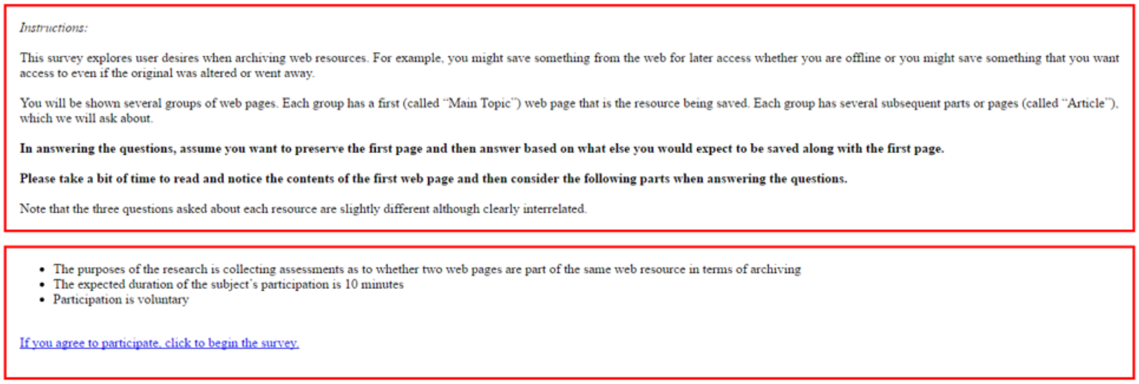
Read More: Seven Principles of Clean Eating

The 7 Best Winter Sangria Recipes
VinePair

20 easy desserts you can make in a cast iron skillet
Domino

If I save the main page, I expect to save this too.	○ Agree	● Disagree
If I have the first page, I would like to have access to this too.	● Agree	○ Disagree
This is part of the resource so I would expect to have access if I save the main page.	● Agree	○ Disagree

Figure 9. Interface for examining and reacting to page pairs. Tabs on left allow participants to switch between the two pages under consideration while agreement with statements is provided below the pages (reprinted from [92]).



Instructions:

This survey explores user desires when archiving web resources. For example, you might save something from the web for later access whether you are offline or you might save something that you want access to even if the original was altered or went away.

You will be shown several groups of web pages. Each group has a first (called "Main Topic") web page that is the resource being saved. Each group has several subsequent parts or pages (called "Article"), which we will ask about.

In answering the questions, assume you want to preserve the first page and then answer based on what else you would expect to be saved along with the first page.

Please take a bit of time to read and notice the contents of the first web page and then consider the following parts when answering the questions.

Note that the three questions asked about each resource are slightly different although clearly interrelated.

- The purposes of the research is collecting assessments as to whether two web pages are part of the same web resource in terms of archiving
- The expected duration of the subject's participation is 10 minutes
- Participation is voluntary

If you agree to participate, click to begin the survey.

Figure 10. Survey Instruction for participants.

Figure 10 shows the instruction page of the user study and explains about the details of the study and what participants need to do. The following instructions explain to the participants at the beginning of the survey:

"*This survey explores user desires when archiving web resources. For example, you might save something from the web for later access whether you are offline or you might save something that you want access to even if the original was altered or went away.*

You will be shown several groups of web pages. Each group has a first (called "Main Topic") web page that is the resource being saved. Each group has several subsequent parts or pages (called "Article"), which we will ask about.

In answering the questions, assume you want to preserve the first page and then answer based on what else you would expect to be saved along with the first page.

Please take a bit of time to read and notice the contents of the first web page and then consider the following parts when answering the questions.

Note that the three questions asked about each resource are slightly different although clearly interrelated.

The purpose of the research is collecting assessments as to whether two web pages are part of the same web resource in terms of archiving."

All data have been saved in our SQL Server database. We preserve session ID, participant's ID, Amazon Mechanical Turk ID which is given to each participant, assessed page Id and each asked question's rate (question 1-3), time of assessment.

4.2 Key Questions to Answer

In the recent study we focus to find the answer to the following questions. For the last question we told the participants in the study that the setting is considered in the personal archive. In the next chapter we will explain the details of another study that considers the institutional archive.

1. Which relationships are important in the pair pages?
2. How similar or different three asked questions are?
3. Is there some type of pattern when their answers are very less or very more?
4. Who is doing the archive? Is it a personal or institutional archive?

4.3 Findings

The results provide insight into the questions raised by our pilot study. We first discuss how reactions to the three different statements provide insight into user expectations and then explore how page relations affect perceptions of resource bounds.

4.3.1 Important Relationships

The following part explains the most important relationships between pair pages. The outcome extracted based on the assessment of the participants in our current user study survey.

- The remaining pages of a story. If a resource has more than one page, the most important and wanted to preserve item is its second page. This interest of archiving drops a bit as it goes to the end of the story pages.
- Images in an image set. When a resource contains a sequence of images, these components have been considered as part of the resource. In this kind of image set, each image has a different URL but all related to the main resource. The results of the study show strong relation between main resource and the remaining image pages.
- Review information of a product. If an Internet resource is about a product, the reviewer opinion page, ratings for an item, more information about the product in Q&A section, and provider information (if there is any) are the most wanted parts of a resource. It is mentionable that all these parts come from various URLs.

- The single page resource itself. If a resource is short enough just including characteristic links and advertisements, the results of our study show the source page itself is the whole thing that participants wanted to archive for later use.

4.3.2 Effects of Differing Content Relations

To understand the effects of different web page relationships on user perceptions of resource bounds, the following discussion only presents the responses to the third statement (S3), which is the explicit assessment of whether the pages are part of the same resource. The data for each of the four resource types is considered independently. Because the navigational distance from the primary page may change perceptions, we present pages at different positions in multi-page stories and image groups separately.

For the resources presenting multi-page stories, the most clear result shown in Figure 11 is that participants consider the later pages of an online story as parts of that resource. Approximately 80% of the participants agree with this assessment. Distance from the first page does have a small influence on this assessment, with fourth pages being less likely to be viewed as part of the same resource than second and third pages (X^2, p<.01). Advertisements are least often viewed as part of the resource (~32%) while links to related content are the most controversial with ~45% agreement that they are part of the same resource.

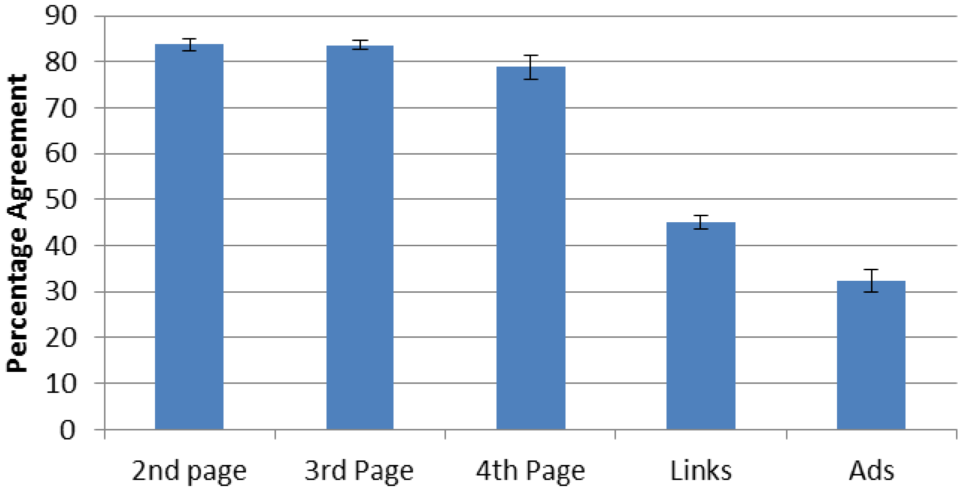


Figure 11. Agreement with S3 for multi-page story resources

Responses for sequences of images (see Figure 12) are similar to those for multi-page stories. While there may be a small drop off in agreement for later images in the image set, results show ~75% agreement that all images in the set are considered as parts of the main resource although they come from different URLs. In short, most participants expect to have access to the rest of the images in the set if they have access to the main one.

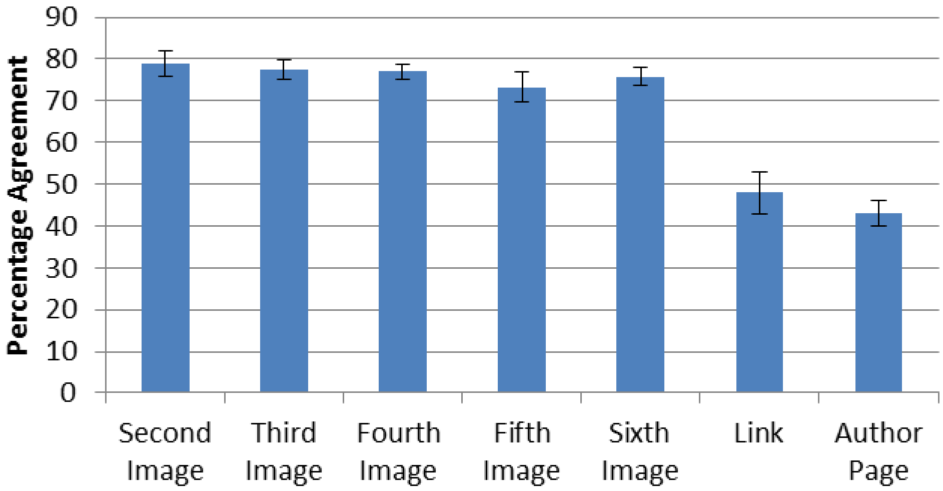


Figure 12. Agreement with S3 for sequence of images resources

About 45% of participants indicate that the linked related content and the page about the photographer/creator are part of the same resource. These results indicate that, as with links to related content found in multi-page stories, there is no clear agreement as to whether these pages should or should not be considered as part of the main resource.

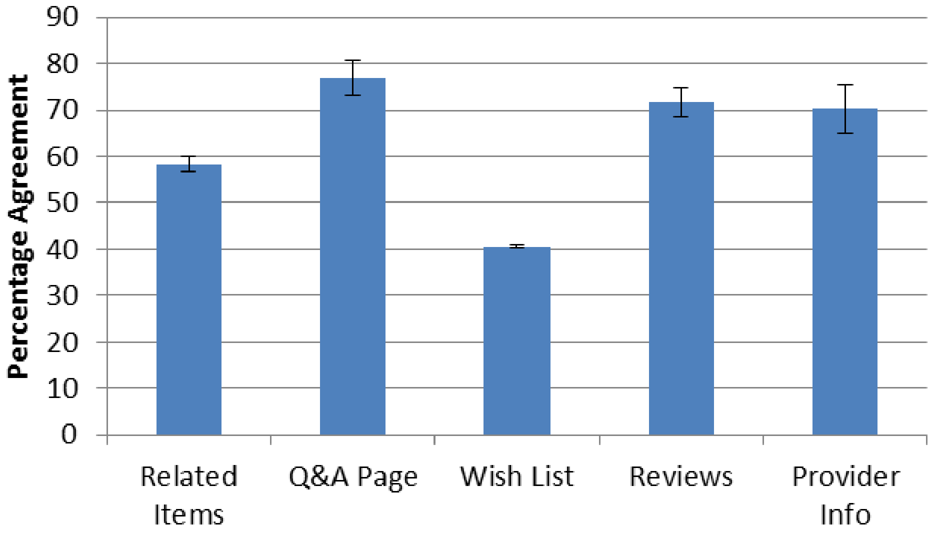


Figure 13. Agreement with S3 for product review resources

Product pages are generally quite complex, combining high-level information about the product with links to related products, more detailed information from the provider/manufacturer, reviews and ratings prior customers or other sources, pages of questions and answers, and wishlists that include the item. Figure 13 shows that 70-75% of respondents tended to view content about the product – the question and answer pages, the reviews and ratings, and the information provided by the manufacturer – as part of the resource identified by the main product page. Fewer respondents (~58%) viewed related items as part of the same resource although they tended to be presented in the same page and only 40% responded that the related wish lists were part of the resource.

When we think of web resources, people often tend to think of completely contained individual web pages. For single page resources (see Figure 14), about 50% of the respondents considered information provided through links to related information or about the author as part of the resource. Consistent with the view in the multi-page stories, only about 35% of respondents considered the advertisements part of the resource.

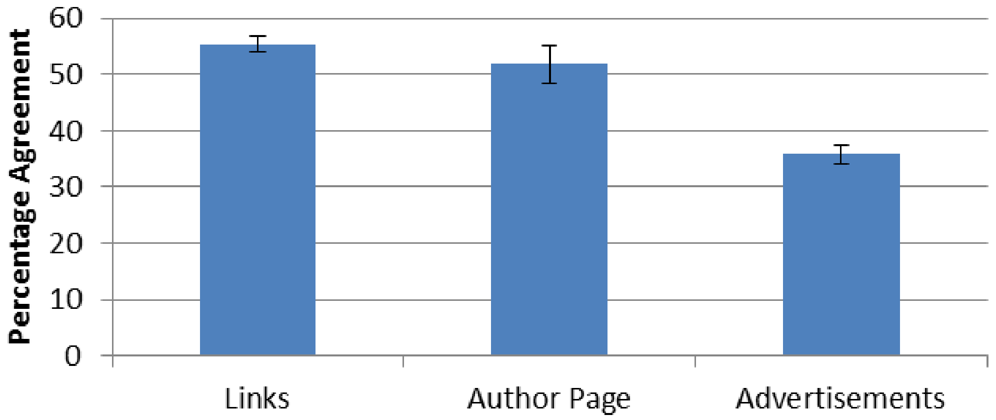


Figure 14. Agreement with S3 for single page resources

Over the four types of page groups considered, the range of agreement that potentially related content was part of the same resource ranged from about 30-35% for advertising that happened to be placed with the main content to 70-80% for strongly-connected content such as the continuation of a story or image set or pages providing more detailed information about products. Other linked content remained near 50% for all types of content.

A closer look at the results for the generic "link" pages connected to the main resources across the three data sets that included them shows that agreement depends to a large degree on the content in the linked page. For example, in one image set group where the primary resource is a food image, one link was to the recipe for the dish and more participants considered this link as part of the resource. In situations where the content in the linked page had a more ambiguous relation to the initial content, users were more ambivalent about inclusion as part of the resource.

4.3.3 Patterns Across the Three Statements

When pair pages were shown to the participants in our survey, the following are three asked questions:

Q1: If I save main page, I expect to save this too.

Q2: If I have the first page, I would like to have access to this too.

Q3: This is part of the resource so I would expect to have access if I save the main page.

The answer options were “agree” or “disagree”. We analyzed the answers to see whether there is a pattern for each question or not. Overall three questions give similar results, and that is what we would expect to get. Q1 gets the most disagreements for all groups.

In other words, the number of agreements for Q1 is slightly less in comparison to the agreements that participants show for Q2 and Q3. This is accurate for all shown pair pages for all groups of resources in the study. Figure 15 shows a sample of answers to three questions and the relation between numerous components of a sample multi-page story group.

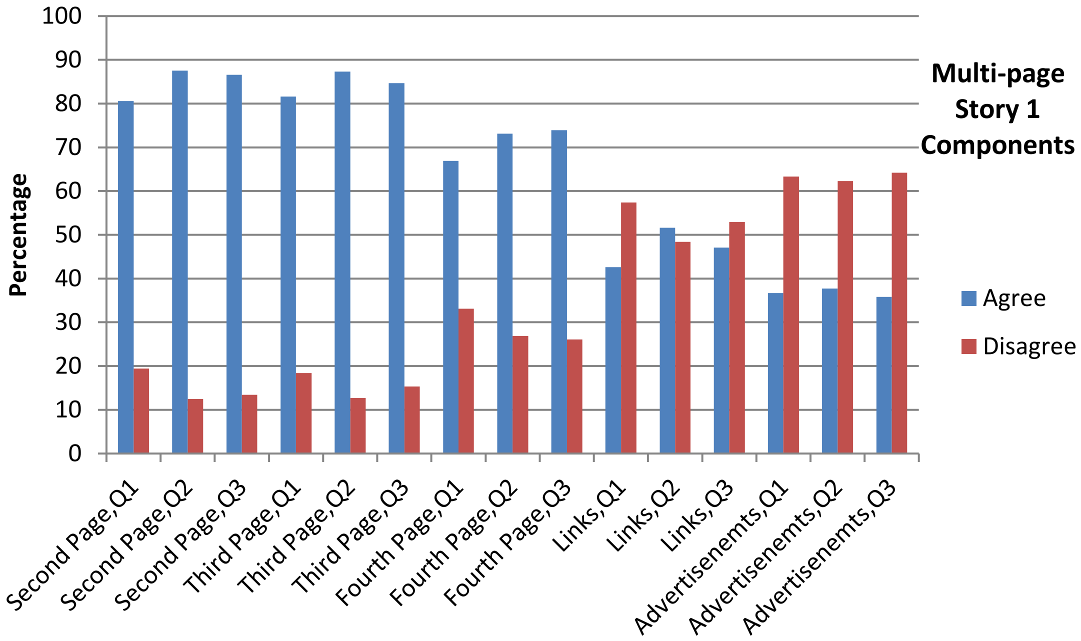


Figure 15. User assessment results for different components of first multi-page story group.

Table 9. Most and least agreeable components in different resources groups

Resource Group	*Most Wanted Components*	*Least Wanted Components*
Multi-page	Second page	Links
Sequence of Images	Second image	Link, author page
Review, Rating and Comment	Reviews, Q&As, provider	Wish list, related item
Single page	Page itself	Link, Ad

Table 9 displays the most and least components of different Internet resources. In multi-page groups, the second page is considered as the obvious part of the resource, while link pages are the least wanted part of the resource. In resources that have review pages, review pages, question and answer pages and the information about the provider of the product or service respectively are expected most to be preserved along with the resource.

For single page resources link and advertisement pages are considered as related components in the conducting survey. The results of the study show that mean (average) agreement for the link pages to be considered as part of the resource is 54.3%. Mean agreement for the expectation of having access to the advertisement pages of a single page resource is 35.95%. Low agreement rate for these component do not nominate them to be part of the resource.

4.4 Summary

As noted before, participants in our study were asked to react to the three statements above (S1, S2, and S3 mentioned in 5.4.3) for a set of page pairs.

Figure 16 compares the results for the three statements across the five multi-page story page groups. Agreements across the three statements follow an observable pattern for pages where the overall level of agreement is high (e.g. later pages.) Participants agree most with the statement that they would like to have access to the second page if they have the first page. This indicates that they value the content in the second page. The participants agree the least with the statement that they expect to have access to the second page if they save the first page, reflecting their mental model of mechanisms for saving web content. The third statement falls in between the other two but is closer to the value-oriented statement (S2) than the expectation-oriented assessment (S1). This indicates that a significant number of participants believe that current mechanisms for saving Internet content do not capture their conception of internet resources.

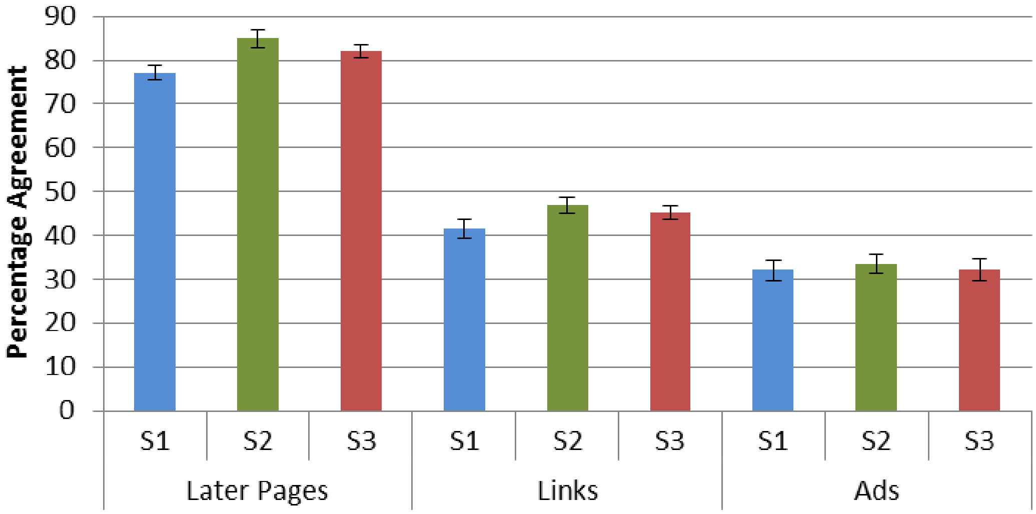


Figure 16. Percentage of respondents that agree with the three statements for multi-page story resources. S1 (expectation when saving) is agreed to the least, S2 (related to value) is agreed to the most, with S3 (part of same resource) is in between.

The other pattern found in participant reactions is that the differences between the reactions to the statements are much lower for less desired content. This indicates that current mechanisms for saving or archiving are not viewed as capturing extraneous content. The patterns found in the reactions to the three statements shown here for multi-page stories are also found in the data for the other resource types.

4.5 Who Does the Archive?

This study asked people about requesting a URL be archived in the context of a personal archive. Thus, they are asked to imagine they want to save the resources for themselves. Would responses be different when thinking about institutional archives and, if so, how? This will be the focus of the next chapter.

4.6 Conclusions

Our prior work found that neither content similarity or URL similarity nor their combination predicted user perception of resource bounds [15]. As a result, this study explores user perceptions for specific relationships between page pairs and variance in responses for value-oriented and expectation-oriented statements regarding the connectedness of pages.

Results of this study show that the relation between perceived content value, resource bounds, and expected system behavior are intertwined. While participant responses to our three statements are very similar, there is a consistent pattern that slightly fewer respondents consider the second page part of the same resource than the number that perceive value in the second page and the number that expect it would be saved is lower still.

The second question explored is how particular relationships between pages affect whether they are considered part of the same resource. When the content on the main and connected pages are parts of a larger composition or set of information, people tend to view the pages as part of the same resource. Similarly, when content on connected pages supports the expected goals of the person visiting the primary page (e.g.

the Q&A pages for products, the recipe for the food image) the perceptions of considering the pages a single resource increase. Incidentally connected content, such as advertisements, tend to not be viewed as part of the resource. More generic links to related content are more idiosyncratic and require domain understanding to predict.

The implication for archiving systems is that techniques for recognizing when content across pages form a composite resource could be applied to better match user desires and expectations. Composite resources result from splitting a single information object (e.g. a text or an image collection) across multiple similar pages, and when a set of typed-components make up a whole (e.g. the various pages for a product on Amazon). While rules can be developed to capture composite resources for particular sites, recognizing these situations more generally would likely require analysis of both the internal structure of page contents and the links between pages. Were it available, navigation data (e.g. which outbound links are most traversed) could also aid in evaluating resource boundaries when available.

5. INSTITUTIONAL ARCHIVE STUDY AND COMPARISON*

This dissertation explores how to determine when additional web pages are part of the same resource as a given page. The preliminary study examined whether relatively simple analyses, such as having similar content, similar URLs, or both would match perceptions [14] for personal archiving scenarios. While there was some correlation, our study found that people often draw their expectations based on a deeper understanding of the relationships between pages and the relationships of the content fragments on those pages.

The second, larger study presented in the prior chapter explored how perceptions of pages with selected relationships, including pages of the same story, images in the same collection, products and reviews/ratings of the product, vary and are related to other content linked to or included in a web page. Users in this study again were presented with a personal archiving scenario as the context for their responses [13]. It was found that pages that included these stronger relationships were more often considered part of the same resource while advertisements were most often considered not part and other linked content was evenly distributed.

This led to the question of whether assumptions about the role of a personal archive were playing a role in the responses [90]. To explore this, we revised the second study to present the decisions in the context of an institutional archive. This chapter

* Reprinted with permission from "How Perceptions of Web Resource Boundaries Differ for Institutional and Personal Archives" by Poursardar, F. and Shipman, F, 2018. IEEE International Conference on Information Reuse and Integration, pp. 126-129, Copyright [2018] by IEEE.

presents the results of this study alongside the results of our second study to examine how perceptions of web resource boundaries differ for institutional and personal archives.

5.1 Study Approach and Participant Instructions

The same survey and resources were used to compare perceptions of web resource boundaries for personal and institutional archives. We previously reported on perceptions of resource boundaries in a personal archiving context [13]. The Amazon Mechanical Turk (MTurk) survey was adapted to describe the same questions about the same resources but in an institutional archiving context.

The introduction to the study and the statements asking for responses from participants were changed in order to be able to compare the results of institutional archive.

In particular, instructions to the participants at the beginning of the survey were:

> *"The current study investigates user expectations when accessing archived web resources. Assume you are making use of an archive provided by an institution tasked with preserving important online resources, such as a digital archive that is part of the Library of Congress. The goal of the archive is to preserve material from the web for later access even if the original was altered or went away.*
>
> *You will be shown several groups of web pages. Each group has a first (called "Main Topic") web page that is the resource being saved by the*

institutional archive. Each group has several subsequent parts or pages (called "Article"), which we will ask about.

In answering the questions, assume the institutional archive has decided it wants to preserve the first page and then answer based on what else you would expect to be saved along with the first page.

Please take a bit of time to read and notice the contents of the first web page and then consider the following parts when answering the questions.

Note that the three questions asked about each resource are slightly different although clearly interrelated."

5.2 Corpus Development and Data Collection

The study courpus consisted of 71 pages included 10 main pages. So 61 pair pages were assessed. The courpus/resource pages and their features are the same as the study of resource boundaries in the context of personal archiving presented in the last chapter. The statements for getting the assessment from participants and introduction changed in order to be able to compare the results of institutional archive.

Consistent with our personal archive study, four categories of page content – multi-page stories, sequence of images, product review and rating, and short single pages – were considered for the institutional archive study. For more detailed information about the resources or the study refer to the previous study in preceding chapter.

5.2.1 Content Presentation and Data Collection

Participants were asked to "agree" or "disagree" with the three following statements S1-S3 for each presented pair page:

S1: If the institutional archive saves the main page, I expect it to save this too.

S2: If the institutional archive has the first page, I would like to have access to this too.

S3: This is part of the resource so I expect the institutional archive would have this if it archives the main page.

- Total assessments are 9018 rows. This is the total number of assessment for all the pages for three statements.
- Three groups (of resource pages) were presented in each task – refer to Table 11 for more details.
- 212 participants assessed the institutional archive survey recruited from Amazon Mechanical Turk. Not all of them completed the survey till the end.
- The study performed on March 1, 2016 till March 15, 2017, and April 3 till April 19, 2017
- Each page assessed between 43 to 68 times, 3 statements asked per page
- Resource pages and their features were divided into 28 groups as shown in the following table (Table 10):

Table 10. Resource Types Considered in the Institutional Archive Study

Type of resource	Number of pages in the group + main resource
Multi-page story resource	6
Sequence of images resource	9
Reviews and ratings resource	7
Short single resource pages	6

For example, each multi-page story resource group has six components or pages that we are getting user perception and assessment on them in order to be able to nominate the key features related to the main or referenced resource in institutional archive senario. In other words, in such a group five pair pages has been shown to each user and got their assessment .

Table 11. Ten Groups and Number of Potential Related Components for Each Group Resource in Institutional Archive Study

Group	Number of page pairs in the group
Multi4 – multi5	5
Img6 – img8	8
Rtg4 – rtg5	6
Sng5 – sng7	5

Above table (see Table 11) shows that we have 10 various groups which each has a number of page pairs. Each time three groups selected randomly through these groups and have been shown to the participants. So the number of assessed page pairs is not exactly equal for each of them. Some of the participants did not complete the whole survey, it means they have quitted somewhere through the task. Some of them contacted me by email and I paid them. If they do not submit the survey completion code, Mechanical Trurk refuses to pay for them. We had 212 participants for the institutional archive data collection study. Total number of assessment for all the pages for all three statements are 9018 records.

For example for the recent 10 groups of page pairs – (including Multi4, multi5, Img6, img7, img8, Rtg4, rtg5, Sng5, sng6, sng7) – we needed to have at least 60 participants to get almost 20 assessment on each pages pair . Each time we select 3 random groups out of above groups, so to get the assessment on 10 groups we need at least 3 participants. It is important to consider that some participants didn't complete their tasks to the end. To make sure that we can get enough assessments we recruited slightly more numbers of participants. Some page pairs got up to 68 assessments, and some got less.

5.3 Multi-page Story Resources

The results of the study for the resourses presenting multi-page story is explained in the follwing section:

1. *Remaining pages of the story.* The most clear result shown in Figure 1 is that participants consider the later pages of an online story as parts of that resource no matter who wants to preserve them. Approximately 80% of the participants agree with this assessment. Distance from the first page does have a small influence on this assessment, with fourth pages being less likely to be viewed as part of the same resource than second and third pages.
2. *Link pages in the main page of the story.* For the link pages, the average agreement (61%) is higher than disagreement, that means participants have considered links as part of the same online multi-page story resource when an institutte wants to preserve it.

3. *Advertisement pages of the story*. The advertisement pages are the most controversial with 45% agreement that they are part of the same resource. The results show that high percentage of participants doesn't consider them as part of the resource.

5.3.1 Institutional and Personal Archive Comparison

For multi-page story resources, the pattern of the boundries are the same; the second page, third page and fourth page consequently are considered parts of the resource.

For multi-page resources the expectation to have access to the links is 45% for personal archive meanwhile expectation to save the same resource components in institutional archive is 61%. Table 12 presents more detail information for the comparison.

Table 12. Different user expectation for personal and institutional archive in multi-page story

Resource Components	Institutional Archive (%)	Personal Archive (%)
Second Page	82.2	83.7
Third Page	80.1	83.6
Fourth Page	74.6	78.9
Links	61.6	45.1
Advertisements	45.4	32.3

Above table (see Table 12) shows the user expectation difference when same resources are considered to be saved as personal archive and institutional archive.The numbers are shown in percentage. Figure 19 represents the same information in a chart. Also Figure 17 and Figure 18 show the agreement with S3 for multi-page story resources for both institutional archive and personal archive.

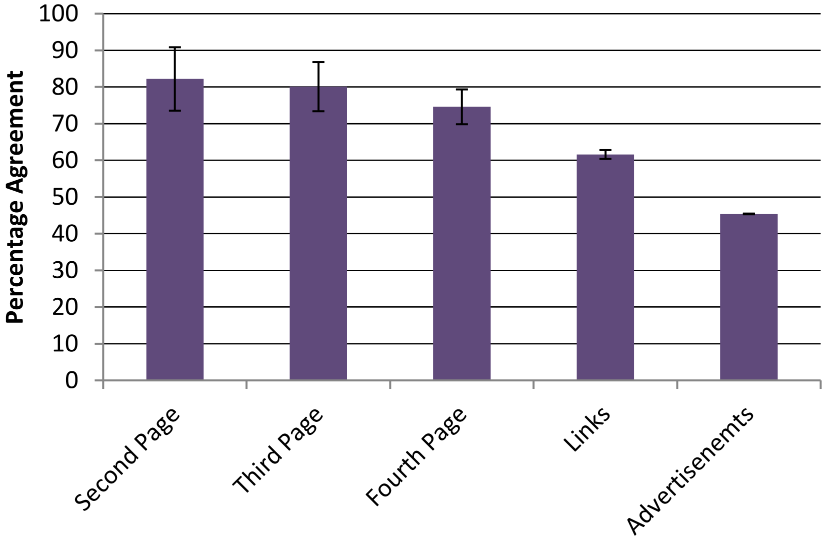


Figure 17. Institutional Archive - Agreement with S3 for multi-page story resources

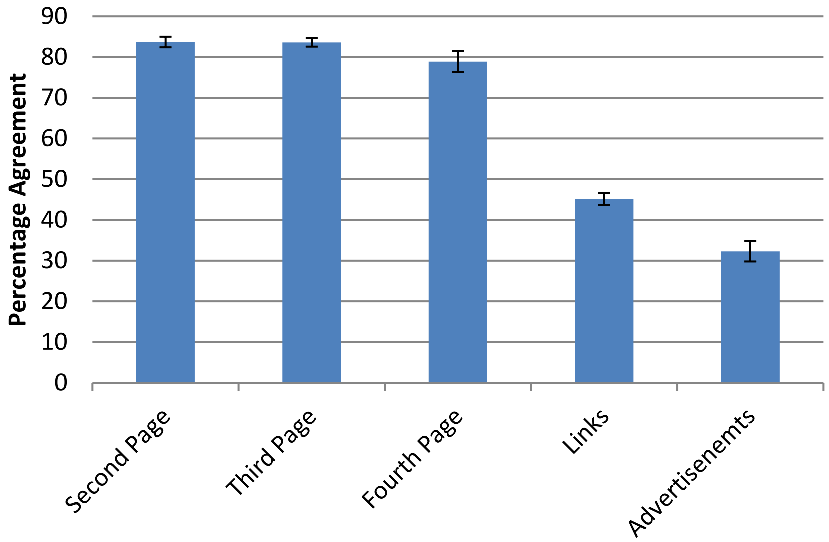


Figure 18. Personal Archive - Agreement with S3 for multi-page story resources

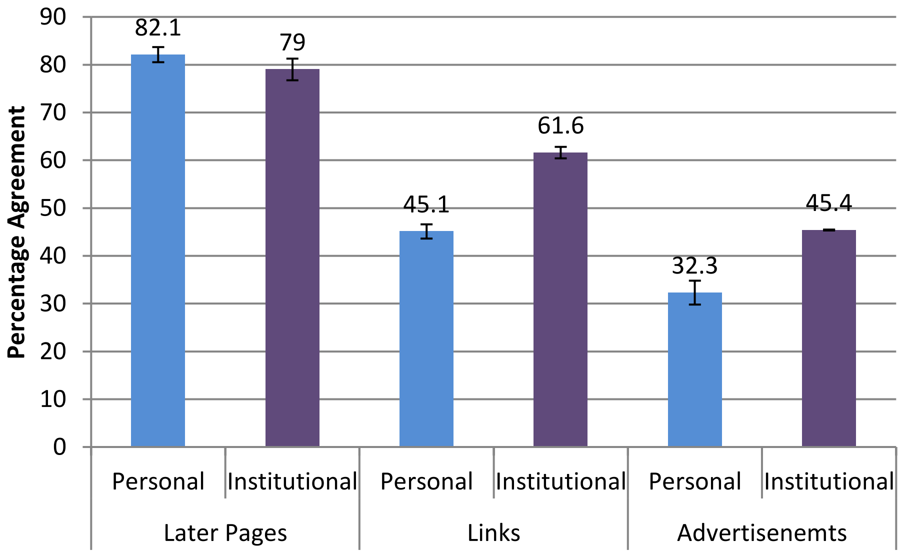


Figure 19. Personal and Institutional Archive Comparison- Agreement with S3 for multi-page story resources

5.4 Sequence of Images Resources

The results of the study for the resourses presenting sequence of images is explained in the follwing section:

1. *Later images of the image set.* It is clear that results of the study show on average approximately 76% agreement that all images in the set are considered as parts of the main resource although they come from different URLs. It means that most of the participants consider later images in a set of images resource as part of that referenced resource for the institutional archive purpose.

2. *Link pages in the main page.* Links to related content have devoted approximately 60% agreement that they are part of the same resource for the institutional archive.

3. *Photographer/creator page.* These pages are referred as author pages in our study and the following charts and tables. With the low rate of user agreement (about 57% - shown in Table 13) on this kind of pages they are the least favorite part of the sequence of images resource in institutional archive.

5.4.1 Institutional and Personal Archive Comparison

Figure 5 and Figure 6 show the agreement with S3 for sequence of images resources for both institutional archive and personal archive. As these figures present all images in the set are considered as the part of the resource no matter who wants to archive them – either it is institutional archive or personal archive. Table 13 shows the user expectation difference when same resources are considered to be saved as personal archive and institutional archive. The numbers are shown in percentage.

About 47% and 58% of participants indicate that the linked related content are part of the same resource consequently for personal and institutional archive. These results for the linked related content indicate that more than half of the participants assess these pages should be considered as part of the main resource for institutional archive. While in personal archive setting, the results for links to related content show no clear agreement as to whether these pages should or should not be considered as part of the main resource.

The pages about the photographer/creator have the same assessment pattern as the links – 43% and 57% for personal and institutional archive studies (see Table 13). It means that participants expect the information about the photographer to be archived for an institution as part of the resource but not worthy enough to be preserved as an personal archive.

Figure 20 represents the personal and institutional archive comparison on agreement with S3 for image sequence resources.

Table 13. Different user expectation for personal and institutional archive in sequence of images

Resource Components	Institutional Archive (%)	Personal Archive (%)
Second Image	81.2	78.9
Third Image	75.4	77.5
Fourth Image	73.9	77
Fifth Image	75.5	73.3
Sixth Image	77.2	75.8
Links	58	47.9
Author Page	57.3	43.1

In the institutiomal archive study survey, it is so interesting that the results turned to be in the similar pattern for our previous study (the personal archive study). For example, the user expectation to archive the second image in the set is high (81%) then it drops a bit for the third and fourth images (75% and 73% respectively) and incrreases again fo the remaining fifth and sixth images of the set to 75 and 77 percentage respectively.

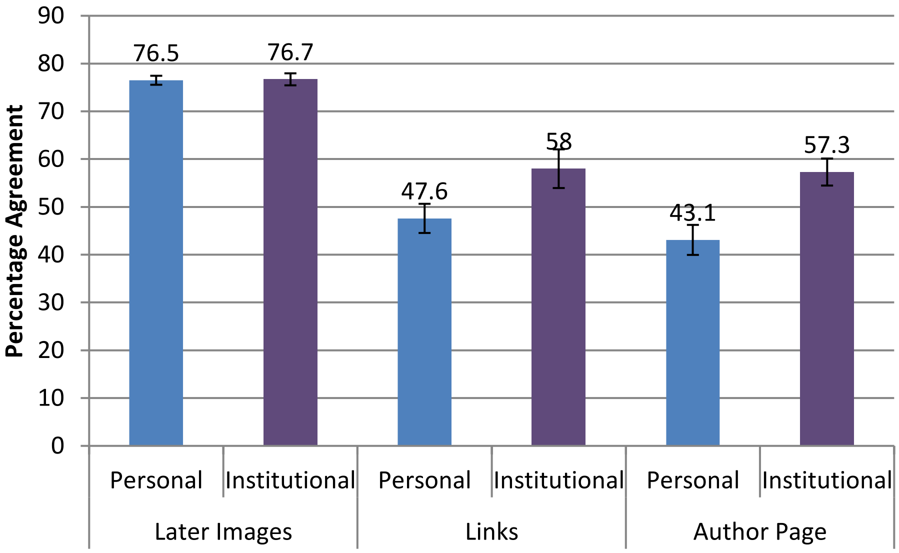


Figure 20. Personal and Institutional Archive Comparison - Agreement with S3 for image sequence resources

5.5 Reviews and Ratings Resources

Review and rating resources considers various type of information about product. These components of the resource are related items, question and answers about the product, information about the manufacturer or provider, review, rating, or comments of the prior customers, and their wish list. The results of the study for institutional archive in Figure 8 show that users perceive all the mentioned components to be part of the resource.

The details on user perception on the boundaries of product pages (review and rating) resources are shown in Table 14. The numbers show the percentage of acceptance of the component as the part of the resource when an institution or an individual save and preserve them for later access.

Table 14. Different user expectation for personal and institutional archive in Reviews and ratings resource

Resource Components	Institutional Archive (%)	Personal Archive (%)
Related Item Pages	64.6	58.3
Q&A Page	73.4	76.9
Wish List	63.3	40.6
Review/Rating Page	65.2	71.6
Provider Information	67.9	70.2

5.5.1 Institutional and Personal Archive Comparison

65 – 73% participants viewed the content related to the product as the resource referenced in the main page of the product in the study for institutional archive. They would expect these components to be preserved by the institute. 70 – 75% respondents acknowledged content related to the product as part of the resource in personal archive setting (see Table 14). The content related to the product are reflected as question and answer pages, review pages and information about the provider/manufacturer.

The respondents tend to expect the institutions to save more components than an individual. The following explain the examples. For this group of resources (review and rating) 63% of participants consider the wish list pages as the part of the resource that an institute should save whilst only 40% of the participants expect the same component as part of the resource when an individual wants to preserve it as his personal archive. Same pattern is true for the related item information which is in the main resource page with consequently 64% and 58% of participants' agreement for institutional and personal archive. Figure 21 represents the personal and institutional archive comparison on agreement with S3 for product review resources.

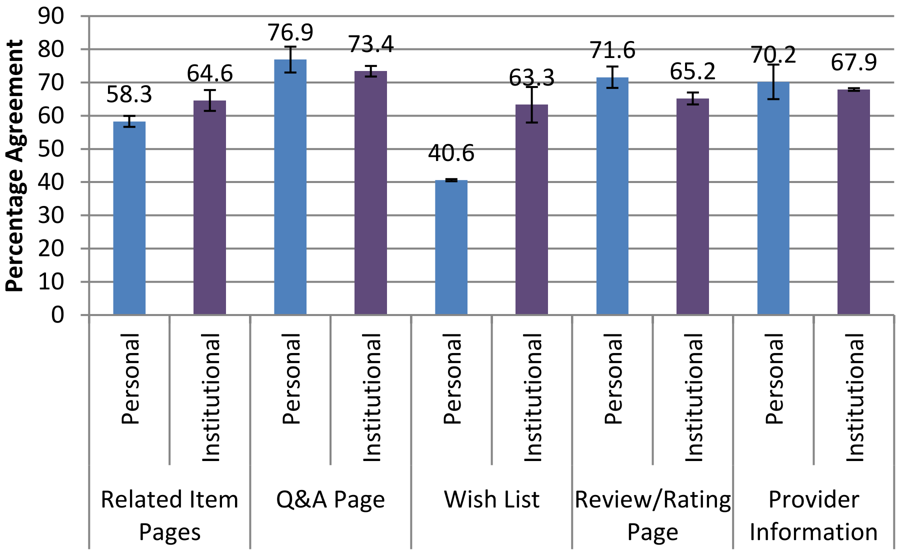


Figure 21. Personal and Institutional Archive Comparison- Agreement with S3 for product review resources

5.6 Single Page Resources

Single resources refer to the short individual traditional single web pages. The following explain about the components and the results in the conducted study. In the study we captured the user expectation when the resource is archived by an institution for later use. Figure 7 presents the results.

1. *Links in the main page.* About 66% of the respondents considered information provided through links to related information as part of the resource. This is the highest expectation in comparison to the other potentially related components to the main resource (see Table 15).

2. *Author of the main page.* 57% of the respondents expect the information about the author of the main page as part of that referenced resource.

3. *Advertisements in the main page.* Approximately 52% of respondents considered the advertisements part of the resource. This component is the least favorite that respondents view as the part of the resource. The outcome reconciles with the multi-page stories.

Table 15. Different user expectation for personal and institutional archive in traditional resources

Resource Components	Institutional Archive (%)	Personal Archive (%)
Links	66.5	55.4
Author Page	57	51.8
Advertisements	52.9	35.8

5.6.1 Institutional and Personal Archive Comparison

Table 15 shows the user expectation difference when same resources are considered to be saved as personal archive and institutional archive. The numbers are shown in percentage. Figure 22 shows the feature agreement comparison for single page resources in personal and institutional archive.

Consistent with the view in personal archive, the pattern of agreement/expectation is the same for institutional archive (refer to Table 15). Links to the related information (main page or resource) has captured the most agreement (about 66%) as part of the resource, then information about the author of the main page with about 55% agreement, and eventually advertisements with about 52% agreement on being as part of the resource.

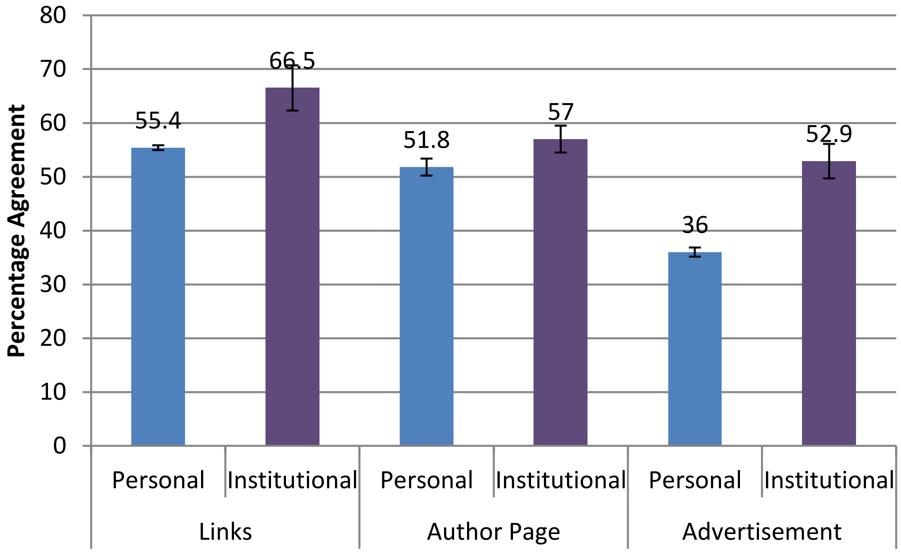


Figure 22. Personal and Institutional Archive Comparison - Agreement with S3 for single page resources

5.7 Conclusion

To sum up, in comparing the results of the conducted personal and institutional studies we can say that the uncertainty increases about the bounds of an Internet resource. People want the institutions to archive more. If the component is not expected to be archive for personal archive it is praiseworthy to be archived for an institute.

Almost all the considered features show the same or a bit increase in user expectation to archive them by an institution. There are some exceptions for fourth image, question and answers pages, rating and review pages, and provider information pages which show the decline of less than 3%. The rating and comment pages which has been dropped to 65.2% in institutional archive setting from 71% in personal archive has the most downfall in about 5%. This drop could be explained that the respondents think/expect the rating or review information of a product may not be as useful and operative as for an institution as in personal archive for a long term access.

6. FEATURE EXTRACTION AND CLASSIFIER DESIGN

Based on the user studies explained in detail in the previous chapters we identified features of the resources and their potential related web pages that are believed to be more valuable when determining the boundaries of internet resources.

What are the implications of these results for software design? Both personal and institutional archives are often limited in what they can capture by available resources. Algorithms that can prioritize the linked content that is most likely to be valued by the user of personal archive or patrons of an institutional archive can make archives' more efficient in the use of available resources. But classification of relationships between primary and linked pages is necessary for the development of such software and this chapter covers them.

In this chapter we explain about the predictive features that have been extracted and used to distinguish whether a component can be considered as part of an Internet resource or not. Later in this chapter we explain how we use the predictive features to design a classifier to classify the resources and their potential components (related web pages to the resource). Then we'll show the result comparison on different classifiers to evaluate how they act upon on selected predictive features.

The chosen features in the composite resources can help us to predict if whether two web pages are part of the same resource or not by performing the automatic classification.

6.1 Predictive Features

The goal is to design classifiers based on some predictive features to automatically distinguish the bounds of an Internet resource. In other words, given two web pages we want to know whether they are part of the same resource or not. For this purpose, we explore the potential for a variety of features to help us in classification. In nominating features we use analysis of both “the internal structure of page contents” and “the links between pages”.

We perform feature extracting phase based on the four different resources types that we have determined previously. These four resource categories are multi-page story, sequence of images, rating and review, and traditional short resources. This section explains about the predictive features for the comparing two pages. We used the features to predict and find out whether the web pages or components are part of the same resource or not. The URLs of the web pages have been used to crawl and extract the needed information. Most of the features present the data that show some kind of relationship between two chosen URLs (web pages). The corpus is the same as the one we have used for our previous user studies - explained in prior chapters. The considered predictive features are the following:

1. *Page_title_similarity*: Two web page titles were compared using Levenshtein distance (LD). It measures the distance between two title strings in our case. It shows the minimum number of edits required to reach from the first web page title to the second one. It could be insertions, deletions or substitutions.

2. *Next_previous_relation*: Given a URL or web page, we crawled it to find out the existence of a specific anchor text in the main frame. We want to know if the web page has some content related or like “next, previous”, “next page”, “previous page”, “1,2,3”, “continued”, “review”, “customer reviews”, “user reviews”, “rate”, “Q & A”, “comment”, etc.
3. *Same_site_address*: this feature defines that how similar are two given URLs. In other words, whether they belong to the same server, or the two comparing pages share the same base URL or belong to the same site. So this value considered in our data.

If the given two URLs (or two given web pages) share the same server or site there is a good chance that they are part of the same resource. For instance, they could be pages of a multi-page story resource, sequence of images, or rating and review resources. If “page number” or similar combinations exist in the page URL, it will be one of the several pages of a story. There is also a chance that they present just some links in the main resource. In order to be able to differentiate this, we need to consider the other features as decision metrics.

4. *Component_Belonging_Average_Agreement*: The average percentage of perception on the component being part of the same resource in each class (multi-story, image-sequence, rating and review resources, and short single page resources).
5. *Component_Belonging_degree_Agreement_percentage* shows the percentage of belonging the component to the same resource based on the user

perception in each resource category. Categories are multi-page story, sequence of images, rating and review, and traditional short resources.

For the following features, we checked some metadata in the web page header:

6. *Number_of_common_Head_base_links*

7. *Author_similarity*

8. *Revision_date_similarity*

9. *Description_similarity*

10. *Title_similarity* defines the title of the document. For instance, for composite resources like sequence of images, if two comparing web pages belong to the same resource, both have the common title of the set as it is shown in the following meta property “og.title”:

<meta property =”og:title” context=”*sequence of images title-title of the set*”>

11. *Number_of_common_Head_rel_links* shows the common components of the web page that two URLs (web pages) share; like common images or icons, etc. Html “link rel” attribute indicates the relationship that the linked resource has to the document from which it’s referenced. In a sequence of image resource it has something similar to the following specification:

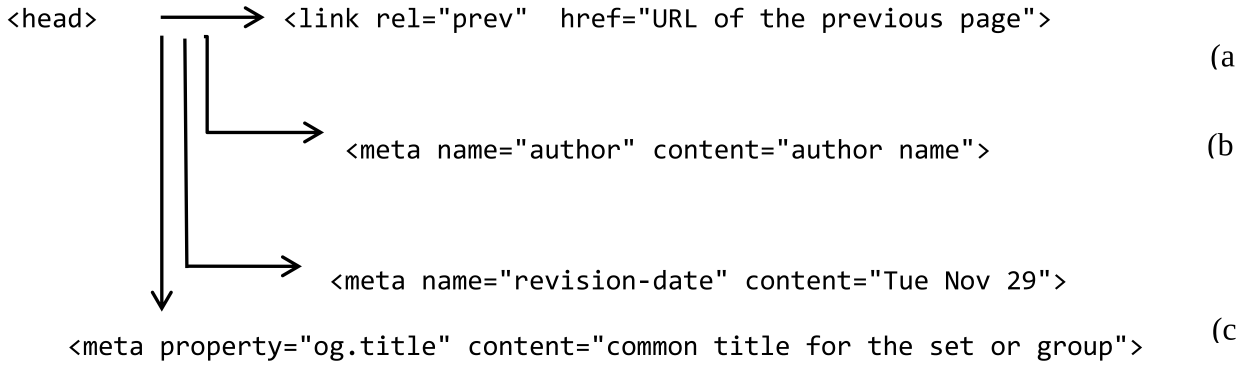
<link rel =”*prev*” href =”*url of the prev*”>

So it presents that “prev” will be part of the current resource and its URL is specified in the “href” tag. The meaning is that each image in the sequence (set) has its own URL and usually base URL is the same and something is added to the end of it to identify the new URL of each image or picture

It is usual for the resources that spread in more than one web page to have some hints in the page header <head> tag. We analyzed inside the <head> tag of the comparing web pages to extract more information on the "description", "title", "author", "revision date"-usually in some types of composite resources it is the same- of the pages, and whether the "link ref href" has the main page's URL or not.

For instance, multi-page story resources main title of the page, author name, and the revision date (if exists) are the same (Figure 23 b and c). Finding similarity of these attributes between two comparing pages is a feature that helps to classify them. Images in the Sequence of images resources share one common title the set in each image page. This feature can be found in meta property =”og.title" (Figure 23 d).

In addition, these composite resource can include "link rel" tags to show the next or previous pages (Figure 23 a).



```
<head>
<link rel="prev"  href="URL of the previous page">
<meta name="author" content="author name">
<meta name="revision-date" content="Tue Nov 29">
<meta property="og.title" content="common title for the set or group">
```

Figure 23. Page header element properties

There were two more feature in our initial deliberations, but they were omitted from further considerations as they didn't demonstrate any changes (improvement or deterioration) in the results. These features are:

Out degree: number of links in the main resource or potential related component

Number of embedded images in the main resource or potential related component

The main focus on the following sections of this chapter will be on using predictive feature to classify the resources.

6.2 Dataset for Classification - Based on Predictive Features

For classification purpose, we have chosen the dataset from the corpus for finding the ground truth in our user studies. The study corpus consists of 24 main/primary pages and 121 pages that were potentially related to from one of the primary pages (shown in Table 16). Four types/categories of web resources were included: multi-page stories, image/photograph sequences, product information, and prototypical single-page resources. For more detailed information about the resources or the study refer to the prior study in Section 5.1. The dataset for classification (121 records) consists of all information extracted of Table 16 pages based on the predictive feature (explained in Section 6.1). The resulting categories and their sizes are shown in Figure 24.

Table 16. Resource Information for classification dataset

Type of Resource	*Number of Groups*	*Number of Components in all Groups minus main page*
Multi-page story resource	5	21
Sequence of images resource	7	44
Reviews and ratings resource	5	24
Short single pages resource	7	32
	24	121

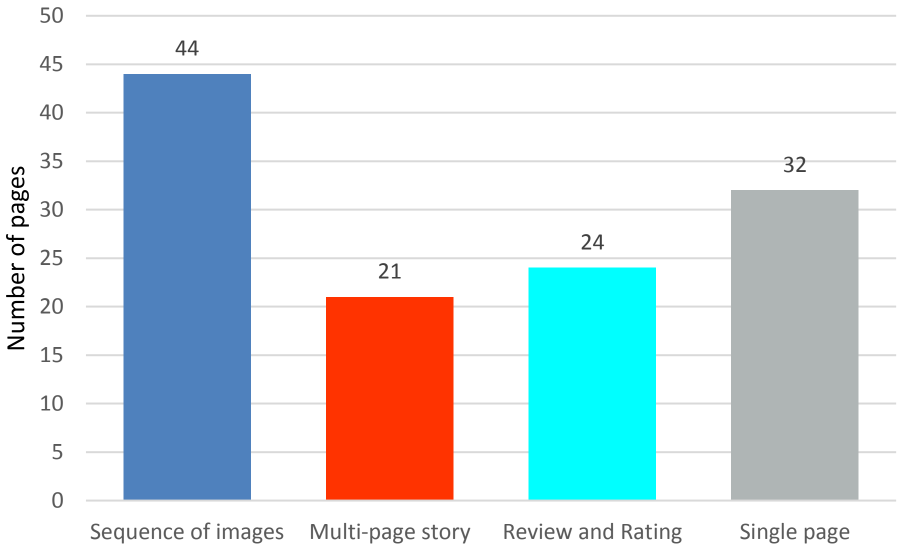


Figure 24. Four different type of resources for classification

Using the explained dataset, we explore the success of different classifiers. Ten-fold cross validation has been used for the evaluation to improve the reliability of the results. In this validation technique, the dataset is randomly divided into ten equal folds, one fold is used as the test set and the remaining as training data set. Then the evaluation repeated ten times and as the result each fold has been used once as the testing set. The 10 results from the folds can then be averaged to produce a single estimation. Ten-fold cross validation is mainly used in systems where the goal is prediction [62], and we wants to estimate how accurately a predictive model will perform. When the dataset is small using cross-validation is a powerful technique to properly estimate model prediction performance [95].

6.3 Classification Algorithms

Now that we have our predictive features and the dataset, the goal is to design classifiers to automatically distinguish the bounds of an Internet resource. In other words, given two web pages we want to know whether they are part of the same resource or not. Later in this chapter we investigate the performance of some different machine learning algorithms and classifiers on our dataset.

For example, a SVM classifier that gets two web pages (two URLs or can be interpreted as two features) and returns yes or no. We train algorithms like SVM to return 70-80% yes, 20-30% no and 15% maybe as the result. Based on these predictive features we want to predict whether “two pages are part of the same resource” or not (Figure 24).

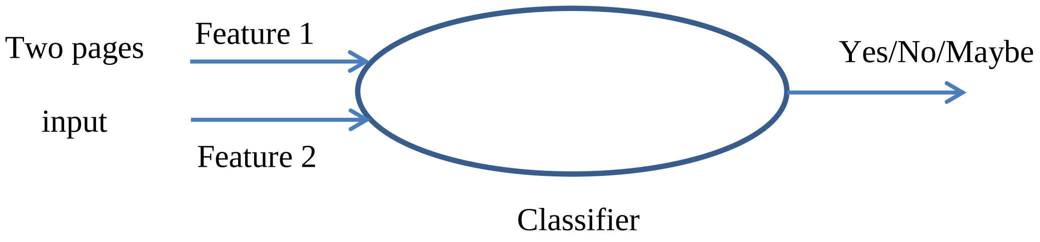


Figure 25. Classifier Design

6.3.1 Evaluation metrics

We choose precision, recall and F-measure as the evaluation measures for our work. Several lines of research [69, 85] have shown that these measures are independent of resource type distributions provided that precision and recall are measured at the same time.

6.4 Single_Resource Classification

As it has been mentioned in the previous sections, our main goal in this research is to be able to distinguish whether two given web pages are part of the same resource or not. To investigate this goal, the results in this section present the classification on “are two pages part of the same resource?”. We call it as “Single_Resource” classification. We have considered to have three classes: “yes”, “no”, and “maybe”. That can be interpreted as “yes, these two pages are part of a resource”, “no, these two pages are not part of a resource”, and “maybe, these two pages could be part of a resource”. Based on the extracted predictive features (in section 6.1) of the web pages in our corpus, we classify and decide on if the two web page can be related to each other.

Table 17. Classifier performance results for Single_Resource classification

	Accuracy	*MAE*	*TP rate*	*FP rate*	*Precision*	*Recall*	*F-Measure*
Bayes Net	84.29%	0.1266	0.843	0.105	0.825	0.843	0.832
Naïve Bayes	76.03	0.1692	0.760	0.103	0.793	0.760	0.774
Logistic	76.03	0.1595	0.760	0.138	0.736	0.760	0.747
Multi-Layer Perception	77.69	0.1603	0.777	0.161	0.736	0.777	0.753
SMO	75.21	0.2865	0.752	0.170	0.704	0.752	0.725
IBK	74.38	0.1781	0.744	0.175	0.738	0.744	0.741
K Star	77.69	0.1573	0.777	0.159	0.754	0.777	0.763
ADA Boost	89.26	0.1343	0.893	0.100	0.888	0.893	0.887
AttributeSelectedClassifier	87.60	0.0949	0.876	0.091	0.883	0.876	0.877
Filtered Classifier	87.60	0.1231	0.876	0.114	0.872	0.876	0.870
Iterative Classifier Optimizer	85.95	0.1204	0.860	0.105	0.859	0.860	0.857
LogiBoost	86.78	0.1119	0.868	0.106	0.858	0.868	0.860
Multiclass Classifier	77.69	0.1933	0.777	0.164	0.717	0.777	0.742
Random Committee	81.82	0.2585	0.818	0.190	0.799	0.818	0.774
Randomizable Filtered	74.38	0.1781	0.744	0.173	0.747	0.744	0.745
Random SubSpace	80.16	0.2938	0.802	0.838	0.769	0.802	0.750
Decision Table	87.60	0.1606	0.876	0.111	0.883	0.876	0.875
JRip	85.95	0.1387	0.860	0.153	0.859	0.860	0.850
PART	84.29	0.1124	0.843	0.128	0.844	0.843	0.841
Hoeffding Tree	76.03	0.1719	0.760	0.105	0.784	0.760	0.771
J48	85.95	0.1278	0.860	0.125	0.852	0.860	0.853
LMT	80.16	0.1517	0.802	0.152	0.776	0.802	0.786
Random Forest	80.99	0.2567	0.810	0.177	0.694	0.810	0.748
ZeroR	57.85	0.3811					

The F-measure of chosen classifiers when given the final 11 features from Section 6.1 to classify the “yes”, “no”, and “maybe” categories for being part of an Internet resource varied between .72 to .89 (shown in Table 17). We used training and testing datasets to the “Single_Resource” classification and performed the evaluation on the designed and some other classifiers. 10 fold cross validation has been used on our dataset to build the training and test sets. The following is the analysis of the classifier results. The values in Table 17 are the weighted average results for each measuring metric for the three classes in Single-Resource classification – “yes”, “no” and “maybe” classes. Measuring metrics are precision, recall and F-measure, MAE, TP, and FP which represent mean absolute error, the number of true positives, and false positives respectively. We have compared 43 different algorithms, and just report the successful algorithms with the best results.

In our study, accuracy for classifier c has been calculated using the below formula (Equation 1):

$$Accuracy(c) = \frac{Items\ classified\ correctly}{All\ items\ classified} \times 100\% \qquad \text{(Equation 1)}$$

We performed our Single_Resource classification with 43 various algorithms that are implemented in the Weka Knowledge Analysis environment for machine learning [106]. We represent the best classification results based on the F-measures from the classifiers types including Bayes, boosting, function, lazy, and decisions trees classifiers. More detailed information about the algorithms of these classifiers can be found in [40, 106].

The results of our investigation for the “Single_Resourse” classification, and the performance metrics for the 23 effective classifiers from our evaluation (among 43) are presented in Table 17. The majority of our classifiers consistently perform with 75 to 89 percent accuracy.

ZeroR classifier performance on “Single_Resource” classification has been included in the last row of Table 17 to just show the “baseline performance on our data”. The ZeroR algorithm also called the Zero Rule is an algorithm that can be used to calculate a baseline of performance for all algorithms on a general dataset. It is the worst result and any algorithm that shows a better performance has some skill on the problem. On a classification algorithm, the ZeroR algorithm will always predict the most abundant category. It is the algorithm you should always run first before all others to develop a baseline. On our dataset, this results in a classification accuracy of 57.85%.

Our study results show that we have nine most effective classifiers with the F-measure greater than .83. These classifiers for the “Single_Resourse” classification - shown in Table 18 below – are ADA Boost, Filtered Classifier, Attribute Selected Classifier, Decision Table, LogiBoost, Iterative Classifier Optimizer, J48, JRip, and Bayes Net respectively. ADA Boost is the best performer for finding parts of an Internet resource classification with 89.26% accuracy and 0.887 F-measure value.

Although our designed SMO classifier performed well with 75.21 % accuracy and F-measure of 0.725, is not among the first best nine classifiers. But its performance nominates it as a good algorithm for “Single_Resource” classification. It implements John Platt's sequential minimal optimization (SMO) algorithm for training a support vector classifier. Also Random committee, Naïve Bayes, K star, Random Forest, and Logistic algorithms are among the algorithms with F-measure metric value higher than .74.

Table 18. The most effective “Single_Resourse” classifiers

		Accuracy	*MAE*	*TP rate*	*FP rate*	*Precision*	*Recall*	*F-Measure*
1	ADA Boost	89.26%	0.1343	0.893	0.100	0.888	0.893	0.887
2	Filtered Classifier	87.60	0.1231	0.876	0.114	0.872	0.876	0.870
3	AttributeSelectedClassifier	87.60	0.0949	0.876	0.091	0.883	0.876	0.877
4	Decision Table	87.60	0.1606	0.876	0.111	0.883	0.876	0.875
5	LogiBoost	86.78	0.1119	0.868	0.106	0.858	0.868	0.860
6	Iterative Classifier Optimizer	85.95	0.1204	0.860	0.105	0.859	0.860	0.857
7	J48	85.95	0.1278	0.860	0.125	0.852	0.860	0.853
8	JRip	85.95	0.1387	0.860	0.153	0.859	0.860	0.850
9	Bayes Net	84.29	0.1266	0.843	0.105	0.825	0.843	0.832

6.4.1 Model Performance Charts and ROCs

In this section, we use Receiver Operating Characteristic (ROC) curves to report the Single_Resource classifiers performance. The ROC graphs display the relative tradeoff between benefits (true positive) rates on the Y axis and the costs (false positive) rate on the X axis. In general ROC curve demonstrates several effects: a) it illustrates true positive rate = tp/(tp+fn) that could be referred to as recall and sensitivity of the algorithm. b) it shows false positive rate = fp/(tn+fp). c) The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test is. d) The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

In our model performance shown in Figure 26 ROC curves show that LogiBoost, Iterative classifier optimizer, and attribute selected classifier offer the best tradeoff between true positive and false positive performance.

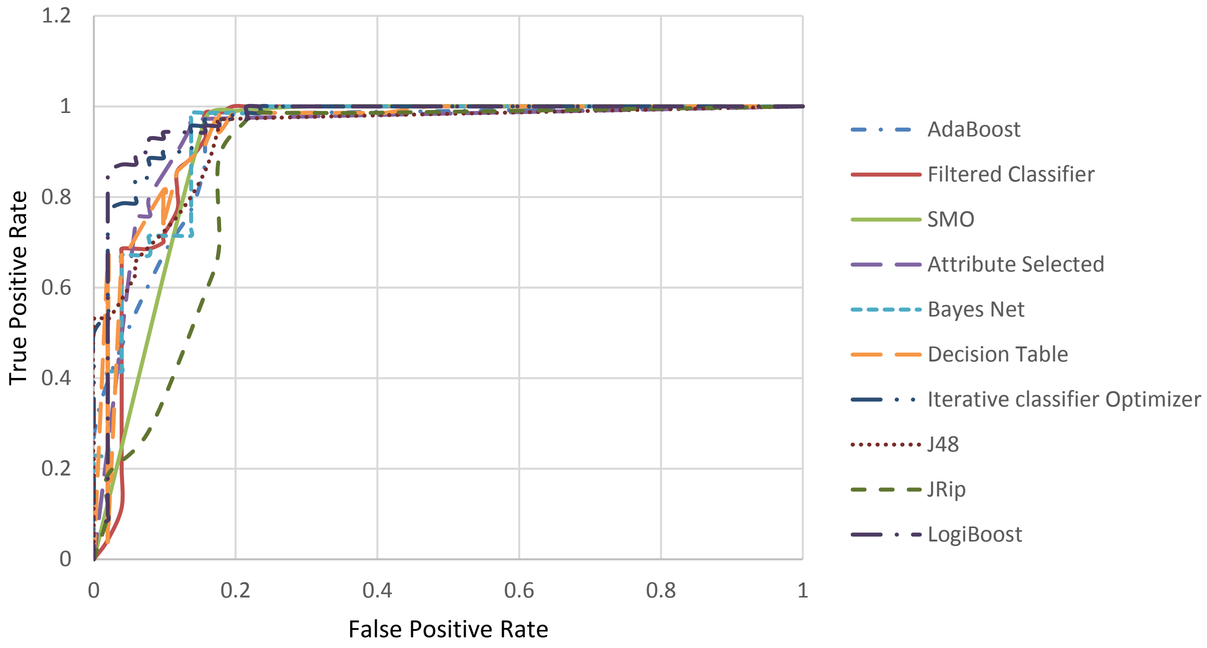


Figure 26. ROC curves to compare model performance for Single_Resource classification

Figure below (Figure 27) shows the model performance chart for our 9 best effective classifiers plus SMO. The color of the curves are based on their threshold performance.

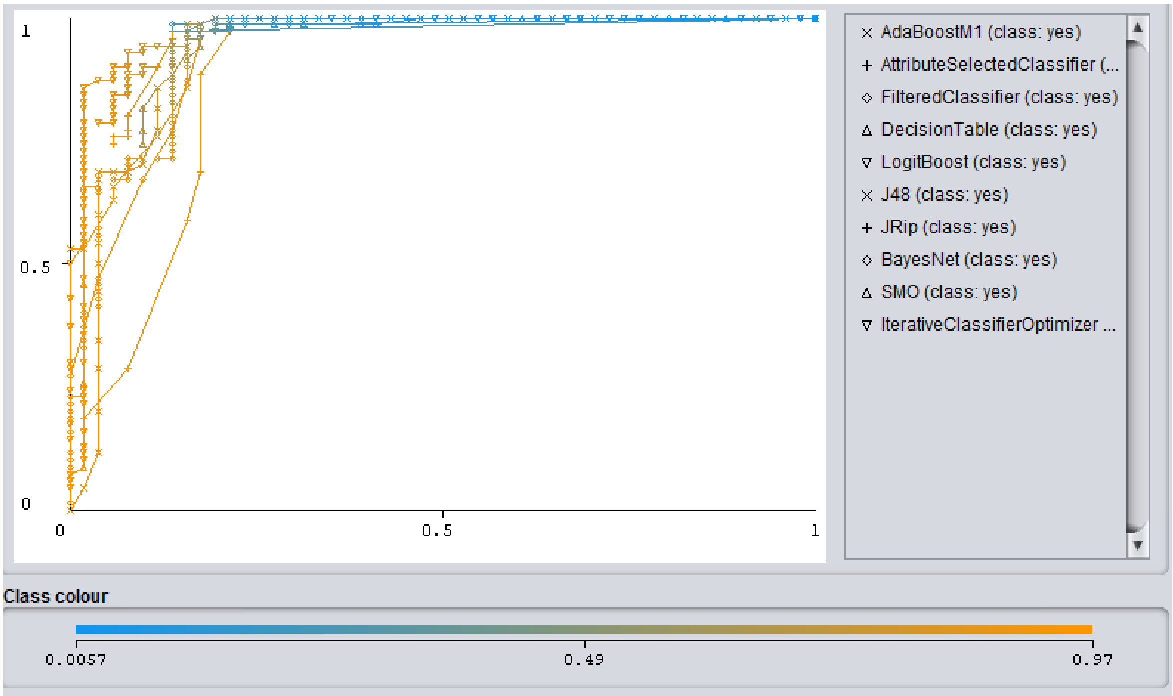


Figure 27. ROC curve to show model threshold performance - Single_Resource classification

6.4.2 Detailed Performance Results

To further investigate the performance of our best classifiers plus SMO algorithm, we illustrate the details of the performing metrics – precision, recall, and F-measure – on each class separately in Table 19. The information on SMO algorithm has been added to the last row for comparison.

Table 19. Detailed performance for best “Single_Resource” classifiers

	Precision			*Recall*			*F-Measure*		
	Maybe	*No*	*Yes*	*Maybe*	*No*	*Yes*	*Maybe*	*No*	*Yes*
ADA Boost	0.714	0.967	0.896	0.556	0.879	0.986	0.625	0.921	0.939
Filtered Classifier	0.714	0.931	0.885	0.556	0.818	0.986	0.625	0.871	0.932
AttributeSelectedClassifier	0.579	1.000	0.905	0.611	0.848	0.957	0.595	0.918	0.931
Decision Table	0.667	1.000	0.883	0.667	0.788	0.971	0.667	0.881	0.925
LogiBoost	0.571	0.935	0.895	0.444	0.879	0.971	0.500	0.906	0.932
Iterative Classifier Optimizer	0.529	0.966	0.893	0.500	0.848	0.959	0.514	0.903	0.924
J48	0.571	0.967	0.870	0.444	0.879	0.957	0.500	0.921	0.912
JRip	0.667	1.000	0.841	0.444	0.818	0.986	0.533	0.900	0.908
Bayes Net	0.462	0.853	0.905	0.333	0.879	0.957	0.387	0.866	0.931
SMO	0.111	0.718	0.849	0.056	0.848	0.886	0.074	0.778	0.867

6.5 Resource_Type Classification

We wanted to know how we can separate our resources based on their types and investigate our classifier algorithms on them. So the results in this section present the classification based on the composite resource types. We call it as “Resource_Type” classification. We have considered to have four classes each for types of our selected Internet resources: “multi-page story resource”, “sequence of images resource”, “rating or review of product resource” and “traditional single page resource”. We used “multi”, “img”, “rtg” and “sng” abbreviation for each class respectively.

The F-measure of chosen classifiers when given the final 11 features from Section 6.1 to classify the categories for resource types varied between .50 to .93 (Table 20). We used training and testing datasets to the “Resource_type” classification and performed the evaluation on the designed and some other classifiers. 10 fold cross validation has been used on our dataset to build the training and test sets. The following is the analysis of the classifier results. The values in Table 20 are the weighted average results for each measuring metric for the four classes in Resource_type classification – “multi”, “img”, “rtg” and “sng” classes. Measuring metrics are precision, recall and F-measure, MAE, TP, and FP which represent mean absolute error, the number of true positives, and false positives respectively. We have compared 40 different algorithms, and just report the successful algorithms with the best results.

Table 20. Classifier performance results for Resource_Type classification

	Accuracy	*MAE*	*TP rate*	*FP rate*	*Precision*	*Recall*	*F-Measure*
IBK	72.73%	0.1448	0727	0.105	0.726	0.727	0.725
K Star	75.51	0.1333	0.785	0.080	0.786	0.785	0.785
AttributeSelectedClassifier	93.38	0.0456	0.934	0.034	0.941	0.934	0.932
Filtered Classifier	88.42	0.081	0.884	0.060	0.898	0.884	0.880
Iterative Classifier Optimizer	87.60	0.1148	0.876	0.061	0.887	0.876	0.873
LogiBoost	87.60	0.1137	0.876	0.061	0.887	0.876	0.873
Random Committee	71.90	0.285	0719	0.136	0.739	0.719	0.685
Random SubSpace	76.03	0.2748	0.760	0.125	0.825	0.760	0.741
Decision Table	91.74	0.1505	0.917	0.047	0.933	0.917	0.916
JRip	86.78	0.092	0.868	0.065	0.883	0.868	0.868
PART	86.78	0.0651	0.868	0.050	0.871	0.868	0.869
J48	89.26	0.0607	0.893	0.049	0.898	0.893	0.891
LMT	82.64	0.1159	0.826	0.080	0.893	0.826	0.821
Random Forest	73.55	0.2924	0.0736	0.123	0.800	0.736	0.715
SMO	52.07	0.3085	0.521	0.216	0.516	0.521	0.491
ZeroR	36.36	0.3648					

The results of our investigation for the “Resource_type” classification, and the performance metrics for the 14 effective classifiers from our evaluation are presented in above Table 20. Last two rows of the table shows the results for SMO and ZeroR algorithms. They have been included for performance comparison with the most effective algorithms and show the baseline. The majority of our classifiers consistently perform with 71 to 93 percent accuracy (shown in Table 20).

We performed our Resource_Type classification with 40 various algorithms that are implemented in the Weka Knowledge Analysis environment for machine learning [106]. We represent the best classification results based on the F-measures from the classifiers types including Bayes, boosting, function, lazy, meta and decisions trees classifiers. More detailed information about the algorithms of these classifiers can be found in [40, 106].

Our study results show that we have fourteen most effective classifiers with the F-measure greater than .83. These classifiers for the “Resource_type” classification - shown in Table 20– are Attribute Selected Classifier, Decision Table, J48, Filtered Classifier, LogiBoost, Iterative Classifier Optimizer, PART, JRip, LMT, Random SubSpace, K start, Random Forest, IBK (K-nearest neighbors classifier), and Random Committee respectively.

In Attribute Selected Classifier dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier. JRip implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by Cohen [19]. Filtered Classifier is a class for running an arbitrary classifier on data that has been passed through an arbitrary filter. Like the classifier, the structure of the filter is based exclusively on the training data and test instances will be processed by the filter without changing their structure. Iterative Classifier Optimizer optimizes the number of iterations of the given iterative classifier using cross-validation or a percentage split evaluation. LMT algorithm is a classifier for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves [65].

Attribute Selected Classifier is the best performer for categorizing based on the resource type with 93.38 % accuracy and 0.932 F-measure value.

Our SMO classifier performed with 52.07% accuracy and F-measure of 0.491, and does not have statistically significant results in comparison to other 14 most effective classifiers in Table 20 for “Resource_type” classification. It implements John

Platt's sequential minimal optimization (SMO) algorithm for training a support vector classifier.

6.5.1 Detailed Performance Results

To further investigate the performance of our best classifiers plus SMO algorithm, we illustrate the details of the performing metrics – precision, recall, and F-measure – on each class separately in below tables (show Table 21 to Table 23). The information on SMO algorithm has been added to the last row for comparison. Algorithms performance are sorted based on their accuracy which are from .93 to .71. First, second and third tables show precision (Table 21), recall (Table 22), and f-measure (Table 23) respectively.

Table 21. Precision performance metric for best Resource_Type classifier algorithms

	Precision			
	Multi	IMG	RTG	SNG
Attribute Selected Classifier	1.000	0.880	1.000	0.941
Decision Table	1.000	0.815	1.000	1.000
J48	0.842	0.840	1.000	0.939
Filtered Classifier	0.929	0.796	1.000	0.941
LogiBoost	0.875	0.792	1.000	0.941
Iterative Classifier Optimizer	0.875	0.792	1.000	0.941
PART	0.727	0.848	0.870	1.000
JRip	1.000	0.796	1.000	0.838
LMT	0.800	0.778	1.000	0.829
Random SubSpace	1.000	0.662	1.000	0.806
K Star	0.900	0.841	0.652	0.735
Random Forest	1.000	0.698	1.000	0.660
IBK	0.783	0.745	0.609	0.750
Random Committee	0.750	0.625	0.950	0.732
SMO	0.667	0.516	0.333	0.556

Table 22. Recall performance metric for best Resource_Type classifier algorithms

	Recall			
	Multi	IMG	RTG	SNG
Attribute Selected Classifier	0.857	1.000	0.792	1.000
Decision Table	0.762	1.000	0.792	1.000
J48	0.762	0.955	0.792	0.969
Filtered Classifier	0.619	0.977	0.792	1.000
LogiBoost	0.667	0.955	0.750	1.000
Iterative Classifier Optimizer	0.667	0.955	0.750	1.000
PART	0.762	0.886	0.833	0.938
JRip	0.762	0.886	0.792	0.969
LMT	0.571	0.955	0.708	0.906
Random SubSpace	0.429	0.977	0.458	0.906
K Star	0.857	0.841	0.625	0.781
Random Forest	0.286	0.841	0.625	0.969
IBK	0.857	0.795	0.583	0.656
Random Committee	0.143	0.795	0.792	0.938
SMO	0.286	0.750	0.167	0.625

Table 23. F-measure performance metric for best Resource_Type classifier algorithms

	F-measure			
	Multi	IMG	RTG	SNG
Attribute Selected Classifier	0.923	0.936	0.884	0.970
Decision Table	0.865	0.898	0.884	1.000
J48	0.800	0.894	0.884	0.954
Filtered Classifier	0.743	0.878	0.884	0.970
LogiBoost	0.757	0.866	0.857	0.970
Iterative Classifier Optimizer	0.757	0.866	0.857	0.970
PART	0.744	0.867	0.851	0.968
JRip	0.865	0.839	0.884	0.899
LMT	0.667	0.857	0.829	0.866
Random SubSpace	0.600	0.789	0.629	0.853
K Star	0.878	0.841	0.638	0.758
Random Forest	0.444	0.763	0.769	0.785
IBK	0.818	0.769	0.596	0.700
Random Committee	0.240	0.700	0.864	0.822
SMO	0.400	0.611	0.222	0.588

6.6 Features Evolution - from Early Work to Predictive

In this section we want to show the continuity from the early work through this chapter and predictive features. This link and comparison will make us able to show whether or not the features identified through the later studies improve our process performance or not. For this purpose, we run the original preliminary features - content similarity, URL similarity, and both - from the preliminary study on the same combinations as the classifiers to understand what their accuracy are.

We have used our most recent corpus (refer to Section 6.1 for more information) for this comparison. We have used this corpus for our complimentary personal and institutional archive as well. The data here compares the effect of including simple techniques for features like URL similarity, content similarity, and both from our preliminary studies.

“Sequence of images” resource groups do not have a lot of text in the web pages. Almost all of them has a different URL, but it is in the same web site as the main page. “Traditional single page” resources have a lot of text to compare and calculate cosine similarity. Two other resource groups – “multi-page story” and “product rating and review” - fit somewhere in between.

In content similarity, we had the simple threshold cutoff (.7) for the cosine similarity in our preliminary study. For the current corpus, the cosine similarity for the main page and each related page in the resource group was calculated. Interestingly, none of the pages in our latest corpus did pass the simple cosine similarity cutoff of .7. So we ignored calculating the accuracy for that. Instead we just used those three features (content similarity, URL similarity, and both) to see if they can be good identifiers on the resource bounds or not. We checked their accuracy using the modern classifications.

Table 24 and Table 25 show the detailed results of applying the primitive features to our most recent corpus performing the evaluation on the same set of classifiers on Sections 6.4 and 6.5. Table 24 shows the results for Single_Resource classification that represents whether pages belong to the same resource.

Table 24. Classifier performance results for primary features for Single_Resource classification

	Accuracy (predictive)	*Accuracy (preliminary)*	*MAE*	*TP rate*	*FP rate*	*Precision*	*Recall*	*F-Measure*
Bayes Net	84.29%	68.59	0.261	0.686	0.200	0.688	0.686	0.686
Naïve Bayes	76.03	65.24	0.301	0.652	0.333	0.561	0.652	0.591
Logistic	76.03	67.76	0.216	0.678	0.277	0.628	0.678	0.649
Multi-Layer Perception	77.69	69.38	0.234	0.693	0.271	0.614	0.693	0.623
SMO	75.21	68.55	0.218	0.685	0.224	0.651	0.685	0.618
IBK	74.38	66.11	0.231	0.661	0.209	0.670	0.661	0.666
K Star	77.69	62.28	0.229	0.653	0.203	0.670	0.653	0.659
ADA Boost	89.26	67.72	0.245	0.677	0.205	-	0.677	-
AttributeSelectedClassifier	87.60	67.27	0.241	0.672	0.207	0.618	0.672	0.608
Filtered Classifier	87.60	68.59	0.294	0.686	0.317	0.601	0.686	0.636
Iterative Classifier Optimizer	85.95	69.38	0.262	0.693	0.264	0.623	0.693	0.628
LogiBoost	86.78	68.80	0.264	0.688	0.272	0.649	0.688	0.647
Multiclass Classifier	77.69	69.38	0.271	0.693	0.292	0.633	0.693	0.614
Random Committee	81.82	66.94	0.339	0.669	0.427	0.611	0.669	0.597
Randomizable Filtered	74.38	57.02	0.290	0.570	0.338	0.553	0.570	0.561
Random SubSpace	80.16	62.80	0.415	0.628	0.510	-	0.628	-
Decision Table	87.60	69.42	0.294	0.694	0.293	0.656	0.694	0.665
JRip	85.95	69.38	0.287	0.693	0.342	0.627	0.693	0.651
PART	84.29	66.90	0.272	0.719	0.326	0.622	0.719	0.658
Hoeffding Tree	76.03	67.27	0.392	0.672	0.324	0.572	0.672	0.583
J48	85.95	66.07	0.291	0.660	0.388	0.554	0.660	0.590
LMT	80.16	67.72	0.294	0.677	0.326	0.584	0.677	0.582
Random Forest	80.99	66.90	0.387	0.669	0.377	-	0.669	-

First column in Table 25 shows accuracy for predictive features - 11 features extracted based on the studies of composite resources. And second column this table represents accuracy of classification for primitive features.

As it is clear in the Table 25 the accuracy have been increased using our new suggested predictive features. The accuracy for applying preliminary features in classification (as shown in Table 25) is between 57 to 69 percent. Meanwhile the accuracy for the predictive feature on the same set of classifiers is between 75 to 89 percent.

Table 25. Accuracy comparison - predictive features vs. preliminary features in Single_Resource classification

	Accuracy (predictive)	*Accuracy (preliminary)*
Bayes Net	84.29%	68.59%
Naïve Bayes	76.03	65.24
Logistic	76.03	67.76
Multi-Layer Perception	77.69	69.38
SMO	75.21	68.55
IBK	74.38	66.11
K Star	77.69	62.28
ADA Boost	89.26	67.72
AttributeSelectedClassifier	87.60	67.27
Filtered Classifier	87.60	68.59
Iterative Classifier Optimizer	85.95	69.38
LogiBoost	86.78	68.80
Multiclass Classifier	77.69	69.38
Random Committee	81.82	66.94
Randomizable Filtered	74.38	57.02
Random SubSpace	80.16	62.80
Decision Table	87.60	69.42
JRip	85.95	69.38
PART	84.29	66.90
Hoeffding Tree	76.03	67.27
J48	85.95	66.07
LMT	80.16	67.72
Random Forest	80.99	66.90

6.6.1 Accuracy Comparison in Resource_Type

Table 24 shows the classification results for including the simple preliminary features - URL similarity, content similarity, and combination of both content and URL similarity- and the comparison with predictive features.

In some classifiers, zero instance could classify to the right class (for at least one class), the number of correctly classified instances are zero. In result we cannot calculate some of the metrics the above table. These cases are shown empty. For example, for ADA Boost classifier the instances for “multi-page story” class and “product rating and review” class are zero, precision and F-measure could not be calculated.

The F-measure of chosen classifiers when given the preliminary features from Chapter 3 to classify the categories for Resource_Types varied between 0.41 to 0.66.

Table 26 We used training and testing datasets to the “Resource_type” classification and performed the evaluation on the same set of classifiers on Sections 6.4 and 6.5. Ten fold cross validation has been used on our dataset to build the training and test sets. Measuring metrics are precision, recall and F-measure, MAE, TP, and FP which represent mean absolute error, the number of true positives, and false positives respectively.

Table 26. Classifiers’ Accuracy comparison - predictive features vs. preliminary features for Resource_Type

	Accuracy (predictive)	*Accuracy (preliminary)*	*MAE*	*TP rate*	*FP rate*	*Precision*	*Recall*	*F-Measure*
SMO	52.07%	50.41%	0.312	0.504	0.235	0.532	0.504	0.489
IBK	72.73	62.80	0.192	0.628	0.156	0.631	0.628	0.628
K Star	75.51	66.94	0.205	0.669	0.122	0.672	0.669	0.669
AttributeSelectedClassifier	93.38	36.36	0.364	0.364	0.364	-	0.364	-
Filtered Classifier	88.42	55.37	0.283	0.554	0.211	-	0.554	-
Iterative Classifier Optimizer	87.60	61.16	0.254	0.612	0.160	0.606	0.612	0.600
LogiBoost	87.60	58.67	0.255	0.587	0.171	0.577	0.587	0.572
Random Committee	71.90	51.23	0.337	0.512	0.246	0.648	0.512	0.460
Random SubSpace	76.03	40.49	0.354	0.405	0.329	-	0.405	-
Decision Table	91.74	52.89	0.304	0.529	0.215	0.504	0.529	0.486
JRip	86.78	62.80	0.246	0.628	0.176	0.651	0.628	0.603
PART	86.78	36.36	0.364	0.364	0.364	-	0.364	-
J48	89.26	36.36	0.364	0.364	0.364	-	0.364	-
LMT	82.64	50.41	0.309	0.504	0.233	0.498	0.504	0.472
Random Forest	73.55	48.76	0.406	0.488	0.275	0.690	0.488	0.417

Table 26 represents the accuracy evaluation results side by side for the preliminary and predictive results together. As show in the table the results are improved a lot when using our new suggested predictive features. The accuracy for applying preliminary features in classification (as shown in Table 26) is between 36 to 66 percent. Meanwhile the accuracy for the predictive feature on the same set of classifiers is mainly between 72 to 93 percent.

6.7 Summary on Learned Lessons and Findings

Given our dataset, we trained the SVM classifier to be able to distinguish that in 70-80% of cases two pages are part of the same resource (classified as “yes” in *Single_Resource* classification), in 20-30% results these two pages are not part of a resource (classified as “no” in *Single_Resource* classification), and about 15% shows that these two pages could be part of a resource (classified as “no” in *Single_Resource* classification).

We used SMO (Sequential Minimal Optimization) because it is a fast training SVM algorithm. In nominating features we use analysis of both the internal structure of page contents and the links between pages. All predictive features for the chosen web pages were included in the dataset. The performance of SMO have shown in the following tables (Table 27 and Table 28).

Table 27. SMO performance for Single_Resource

	Accuracy	*MAE*	*TP rate*	*FP rate*	*Precision*	*Recall*	*F-Measure*
SMO	75.21%	0.2865	0.752	0.170	0.704	0.752	0.725

Table 28. SMO detailed performance for Single_Resource

	Precision			*Recall*			*F-Measure*		
	Maybe	*No*	*Yes*	*Maybe*	*No*	*Yes*	*Maybe*	*No*	*Yes*
SMO	0.111	0.718	0.849	0.056	0.848	0.886	0.074	0.778	0.867

Our study results showed that we have nine most effective classifiers with the F-measure greater than .83 for the “Single_Resourse” classification - shown in Table 18. ADA Boost is the best performer for finding parts of an Internet resource classification with 89.26% accuracy and 0.887 F-measure value. Although our SMO classifier performed well with 75.21 % accuracy and F-measure of 0.725, it is not among the first best nine classifiers. But its performance still nominates it as a good algorithm for “Single_Resource” classification.

Our model performance ROC curves show that LogiBoost, Iterative classifier optimizer, and attribute selected classifier offer the best tradeoff between true positive and false positive performance (refer to Figure 26 for more information).

In the second part of this chapter, we considered to study on if we separate the composite Internet resources - four type of resources that we have considered in this study- based on their distinguished types how our classifiers can perform and what will be the results. Our study results show that we have fourteen most effective classifiers with the F-measure greater than .83. These classifiers for the “Resource_type” classification are Attribute Selected Classifier, Decision Table, J48, Filtered Classifier, LogiBoost, Iterative Classifier Optimizer, PART, JRip, LMT, Random SubSpace, K start, Random Forest, IBK (K-nearest neighbors classifier), and Random Committee respectively- refer to Table 20. SMO classifier performed with 52.07% accuracy and F-measure of 0.491, and does not have statistically significant results in comparison to other 14 most effective classifiers for “Resource_type” classification.

7. CONCLUSION

In this dissertation, we have introduced and developed algorithms and techniques to promote web archiving systems in identifying the bounds of an Internet resource. We introduced algorithms that automatically distinguish whether two web pages are part of the same resource or not. There are three objectives that we have discussed throughout this dissertation:

First, we studied user perception on the bounds of the resources. We conducted several major user studies to be able to distinguish and differ the predicting features.

The results of our first primary study showed that the features that make a difference in whether people expect to have access to content are more nuanced than simply having similar content or a similar URL. These results indicate that content similarity is likely a viable feature for systems when deciding the bounds of a resource.

Based on the results we felt the need of more investigation to help design techniques to automatically identify the bounds of a resource. So our later studies explore user perceptions on complex resources for specific relationships between page pairs and variance in responses for value-oriented and expectation-oriented statements regarding the connectedness of pages.

Results of this study show that the relation between perceived content value, resource bounds, and expected system behavior are intertwined. While participant responses to our three statements are very similar, there is a consistent pattern that slightly fewer respondents consider the second page part of the same resource than the number that perceive value in the second page and the number that expect it would be saved is lower still.

The second question explored is how particular relationships between pages affect whether they are considered part of the same resource. When the content on the main and connected pages are parts of a larger composition or set of information, people tend to view the pages as part of the same resource. Similarly, when content on connected pages supports the expected goals of the person visiting the primary page (e.g. the Q&A pages for products, the recipe for the food image) the perceptions of considering the pages a single resource increase. Incidentally connected content, such as advertisements, tend to not be viewed as part of the resource. More generic links to related content are more idiosyncratic and require domain understanding to predict.

The implication for archiving systems is that techniques for recognizing when content across pages form a composite resource could be applied to better match user desires and expectations. Composite resources result from splitting a single information object (e.g. a text or an image collection) across multiple similar pages, and when a set of typed-components make up a whole (e.g. the various pages for a product on Amazon). While rules can be developed to capture composite resources for particular sites, recognizing these situations more generally would likely require analysis of both the internal structure of page contents and the links between pages. This could be supported by patterns in usage data when available.

Second, we extended our studies and wanted to inspect the user perception in institutional archive. It is very interesting to investigate that if “who does the archive” can make an impact on the boundaries of the Internet resources or not.

Consistent with our previous study on personal archiving, the primary-page content in the study comes from multi-page stories, multi-image collections, product pages with reviews and ratings on separate pages, and short single page writings. Participants were asked to assume the institutional archive wants to preserve the primary page and then answer what else they would expect to be saved along with the primary page.

This study extends our initial study in the context of personal archiving and confirms similar patterns of results for institutional archives. When a single story or sequence of images is spread across multiple web pages, there is general agreement that these pages are part of single web resource. The largest differences in perceptions came from the least-related content pairs presented. Advertising, wishlists, and generic links were all much more likely to be considered part of the same resource for institutional archives than they were for personal archives. Perhaps it is that the uncertainty of future uses in institutional settings increases the uncertainty about the bounds of an Internet resource. In any case, these results show that people want institutional archives to capture more than similar personal archives.

The only significant drop from the personal archiving scenario to the institutional archiving scenario was for product rating and comment pages (about 5%.) It would be interesting to explore this result in follow-on studies to determine whether assumptions about the use of the resource or the fluidity of the content (or neither) are contributing to this result.

Finally, across both settings, there was never agreement by more than 82% that two pages were part of the same resource. Perhaps this indicates that some respondents (~18%) view each web page as a unique resource, either because they experience it that way or because they want to avoid more ambiguous rules for the bounds of resources. Also, there was never lower than 32% that believed the two pages presented were part of the same resource. This could indicate that respondents have a different notion of what it means to be a web resource (conflating resource and provider) or that their responses are meant to indicate a desire that everything be archived.

Finally, all of our studies confirmed that perceptions of resource boundaries are consistent across the considered applications. What are the implications of these results for software design? Both personal and institutional archives are often limited in what they can capture by available resources. Algorithms that can prioritize the linked content that is most likely to be valued by the user of personal archive or patrons of an institutional archive can make archives' more efficient in the use of resources. Classification of relationships between primary and linked pages is necessary for the development of such software. We used the predictive features of the web resource boundaries and designed algorithms to classify them. In this process the recommended algorithms can distinguish whether two input the web pages are part of the same resource.

7.1 Future Work

As indicated by the variance in human perceptions, the variety of contexts and features explored, and the results of the designed classifiers, reasoning about resource boundaries is challenging. This is not just a problem for archiving systems. Knowing what is and is not part of a resource is also important when indexing resources for search, recommendation, and visualization. While those contexts may well elicit alternative perceptions of resource bounds than were found in the archiving contexts explored here, it is also likely that some of the same features will be valuable in such contexts. One difference is that the costs associated with false positives and false negatives may well be different when building such systems. Future work exploring human perceptions in such contexts and the ability of techniques for determining bounds to meet the particular needs of those contexts is a goal of future work.

REFERENCES

[1] Abrams, D., Baecker, R., and Chignell, M., "Information archiving with bookmarks: personal Web space construction and organization," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1998, pp. 41-48.

[2] Abrams, S. L. and Seaman, D., "Towards a global digital format registry," 2003.

[3] Ashman, H., "Electronic document addressing: dealing with change," *ACM Computing Surveys (CSUR),* vol. 32, pp. 201-212, 2000.

[4] Baeza-Yates, R., Pereira, Á., and Ziviani, N., "Genealogical trees on the web: a search engine user perspective," in *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 367-376.

[5] Bälter, O., "Keystroke level analysis of email message organization," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 2000, pp. 105-112.

[6] Bar-Yossef, Z., Broder, A. Z., Kumar, R., and Tomkins, A., "Sic transit gloria telae: towards an understanding of the web's decay," in *Proceedings of the 13th international conference on World Wide Web*, 2004, pp. 328-337.

[7] Barreau, D. and Nardi, B. A., "Finding and reminding: file organization from the desktop," *ACM SigChi Bulletin*, vol. 27, pp. 39-43, 1995.

[8] Barreau, D. K., "Context as a factor in personal information management systems," *Journal of the American Society for Information Science*, vol. 46, p. 327, 1995.

[9] Bellotti, V., Dalal, B., Good, N., Flynn, P., Bobrow, D. G., and Ducheneaut, N., "What a to-do: studies of task management towards the design of a personal task list manager," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 735-742.

[10] Bellotti, V., Ducheneaut, N., Howard, M., and Smith, I., "Taking email to task: the design and evaluation of a task management centered email tool," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003, pp. 345-352.

[11] Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., and Grinter, R. E., "Quality versus quantity: E-mail-centric task management and its relation with overload," *Human-computer interaction,* vol. 20, pp. 89-138, 2005.

[12] Bellotti, V., Ducheneaut, N., Howard, M., Smith, I., and Neuwirth, C., "Innovation in extremis: evolving an application for the critical work of email and information management," in *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, 2002, pp. 181-192.

[13] Blandford, A. E. and Green, T. R., "Group and individual time management tools: what you get is not what you need," *Personal and Ubiquitous Computing,* vol. 5, pp. 213-230, 2001.

[14] Boardman, R. and Sasse, M. A., "Stuff goes into the computer and doesn't come out: a cross-tool study of personal information management," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 583-590.

[15] Bondarenko, O. and Janssen, R., "Documents at hand: Learning from paper to improve digital technologies," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2005, pp. 121-130.

[16] Bruce, H., Jones, W., and Dumais, S., "Information Behaviour That Keeps Found Things Found," *Information Research: An International Electronic Journal,* vol. 10, p. n1, 2004.

[17] Brunelle, J. F., Kelly, M., SalahEldeen, H., Weigle, M. C., and Nelson, M. L., "Not all mementos are created equal: Measuring the impact of missing resources," *International Journal on Digital Libraries,* vol. 16, pp. 283-301, 2015.

[18] Carroll, J. M., "Creative names for personal files in an interactive computing environment," *International Journal of Man-Machine Studies,* vol. 16, pp. 405-438, 1982.

[19] Cohen, W. W., "Fast effective rule induction," in *Machine Learning Proceedings 1995,* ed: Elsevier, 1995, pp. 115-123.

[20] Cornell, U., "Digital Preservation Management: Implementing Short-Term Strategies for Long-Term Solutions," *online tutorial developed for the Digital Preservation Management workshop,* vol. section 4B, 2003.

[21] Crow, R., "The case for institutional repositories: A SPARC position paper, 2002," *Washington, DC, SPARC: http://www. arl. org/sparc/bm~ doc/ir_final_release_102. pdf11,* 2002.

[22] Cunningham, A., "Waiting for the ghost train: Strategies for managing electronic personal records before it is too late," *Archival Issues,* pp. 55-64, 1999.

[23] Cunningham, S. J. and Masoodian, M., "Identifying personal photo digital library features," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 2007, pp. 400-401.

[24] Dalal, Z., Dash, S., Dave, P., Francisco-Revilla, L., Furuta, R., Karadkar, U., *et al.*, "Managing distributed collections: evaluating web page changes, movement, and replacement," in *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, 2004, pp. 160-168.

[25] Darlington, J., "Pronom-a practical online compendium of file formats," *RLG DigiNews,* vol. 7, 2003.

[26] Davis, H. C., "Hypertext link integrity," *ACM Computing Surveys (CSUR),* vol. 31, p. 28, 1999.

[27] Davis, S., "Electronic Records Planning in" Collecting" Repositories," *The American Archivist,* vol. 71, pp. 167-189, 2008.

[28] Day, M., "Preserving the fabric of our lives: a survey of Web preservation initiatives," in *Research and Advanced Technology for Digital Libraries*, ed: Springer, 2003, pp. 461-472.

[29] Decman, M., "Problems of Long-Term Preservation of Web Pages," *Knjiznica,* vol. 55, pp. 193-208, 2011 2011.

[30] Dix, A. and Marshall, J., "At the right time: when to sort web history and bookmarks," ed: Lawrence Erlbaum Associates, 2003, pp. 758-762.

[31] Ducheneaut, N. and Bellotti, V., "E-mail as habitat: an exploration of embedded personal information management," *interactions,* vol. 8, pp. 30-38, 2001.

[32] Elsweiler, D., Ruthven, I., and Jones, C., "Towards memory supporting personal information management tools," *Journal of the American Society for Information Science and Technology,* vol. 58, pp. 924-946, 2007.

[33] Fetterly, D., Manasse, M., Najork, M., and Wiener, J., "A large-scale study of the evolution of web pages," in *Proceedings of the 12th international conference on World Wide Web*, 2003, pp. 669-678.

[34] Francisco-Revilla, L., Shipman, F., Furuta, R., Karadkar, U., and Arora, A., "Managing change on the web," in *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, 2001, pp. 67-76.

[35] Francisco-Revilla, L., Shipman III, F. M., Furuta, R., Karadkar, U., and Arora, A., "Perception of content, structure, and presentation changes in Web-based hypertext," in *Proceedings of the 12th ACM conference on Hypertext and Hypermedia*, 2001, pp. 205-214.

[36] Furuta, R., Shipman III, F. M., Marshall, C. C., Brenner, D., and Hsieh, H.-w., "Hypertext paths and the World-Wide Web: experiences with Walden's Paths," in *Proceedings of the eighth ACM conference on Hypertext*, 1997, pp. 167-176.

[37] Garde-Hansen, J., "MyMemories?: Personal digital archive fever and Facebook," in *Save as… Digital memories*, ed: Springer, 2009, pp. 135-150.

[38] Gottlieb, L. and Dilevko, J., "User preferences in the classification of electronic bookmarks: Implications for a shared system," *Journal of the Association for Information Science and Technology*, vol. 52, pp. 517-535, 2001.

[39] Gschwind, R., Rosenthaler, L., and Büchel, R., "Digitization and Long Term Archival of Photographic Collections: Recommendations of the Swiss Federal Office for Civil Protection, Section Protection of Cultural Property," in *Archiving Conference*, 2004, pp. 11-17.

[40] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H., "The WEKA data mining software: an update," *ACM SIGKDD explorations newsletter*, vol. 11, pp. 10-18, 2009.

[41] Harper, R., Lindley, S., Thereska, E., Banks, R., Gosset, P., Smyth, G., *et al.*, "What is a File?," in *Proceedings of the 2013 conference on Computer supported cooperative work*, 2013, pp. 1125-1136.

[42] Henderson, S., "How do people organize their desktops?," in *CHI'04 extended abstracts on Human factors in computing systems*, 2004, pp. 1047-1048.

[43] Hitchcock, S., Brody, T., Hey, J., and Carr, L., "Preservation for institutional repositories: practical and invisible," 2005.

[44] Hockx-Yu, H., "The past issue of the web," in *Proceedings of the 3rd International Web Science Conference,* 2011, p. 12.

[45] Huurdeman, H. C., Kamps, J., Samar, T., de Vries, A. P., Ben-David, A., and Rogers, R. A., "Lost but not forgotten: finding pages on the unarchived web," *International Journal on Digital Libraries,* vol. 16, pp. 247-265, 2015.

[46] JHOVE, J., "Harvard Object Validation Environment," *Information available at: http://hul. harvard. edu/jhove/index. html*.

[47] John, J. L., Rowlands, I., Williams, P., and Dean, K., "Digital lives: personal digital archives for the 21st century: an initial synthesis," *Version 0.2 (March 3, 2010),* 2010.

[48] Jones, W., "The future of personal information management, part I: our information, always and forever," *Synthesis lectures on information concepts, retrieval, and services,* vol. 4, pp. 1-125, 2012.

[49] Jones, W., *Keeping Found Things Found: The Study and Practice of Personal Information Management: The Study and Practice of Personal Information Management*: Morgan Kaufmann, 2010.

[50] Jones, W., Bruce, H., and Dumais, S., "Keeping found things found on the web," presented at the Proceedings of the tenth international conference on Information and knowledge management, Atlanta, Georgia, USA, 2001.

[51] Jones, W., Phuwanartnurak, A. J., Gill, R., and Bruce, H., "Don't take my folders away!: organizing personal information to get ghings done," in *CHI'05 extended abstracts on Human factors in computing systems*, 2005, pp. 1505-1508.

[52] Kahle, B., "Preserving the internet," *Scientific American,* vol. 276, pp. 82-83, 1997.

[53] Karger, D. R. and Jones, W., "Data unification in personal information management," *Communications of the ACM,* vol. 49, pp. 77-82, 2006.

[54] Kaye, J. J., Vertesi, J., Avery, S., Dafoe, A., David, S., Onaga, L., *et al.*, "To have and to hold: exploring the personal archive," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006, pp. 275-284.

[55] Kim, J., "Motivating and impeding factors affecting faculty contribution to institutional repositories," *Journal of digital information,* vol. 8, 2007.

[56] Kim, P., Podlaseck, M., and Pingali, G., "Personal chronicling tools for enhancing information archival and collaboration in enterprises," in *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, 2004, pp. 56-65.

[57] Klein, M., "Using the web infrastructure for real time recovery of missing web pages " Doctor of Philosophy, Old Dominion University, 2011.

[58] Klein, M. and Nelson, M., "Moved but not gone: an evaluation of real-time methods for discovering replacement web pages," *International Journal on Digital Libraries,* vol. 14, pp. 17-38, 2014/04/01 2014.

[59] Klein, M. and Nelson, M. L., "Evaluating methods to rediscover missing web pages from the web infrastructure," in *Proceedings of the 10th annual joint conference on Digital libraries*, 2010, pp. 59-68.

[60] Klein, M., Ware, J., and Nelson, M. L., "Rediscovering missing web pages using link neighborhood lexical signatures," in *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, 2011, pp. 137-140.

[61] Koehler, W., "Web page change and persistence—A four-year longitudinal study," *Journal of the American Society for Information Science and Technology,* vol. 53, pp. 162-171, 2002.

[62] Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, 1995, pp. 1137-1145.

[63] Kriegsman, S. and Mandell, L., "Digital Archiving Without Preservation is Just Storage: Education is the First Step to Achieving Preservation Goals," in *Archiving Conference*, 2004, pp. 32-35.

[64] LaBarca, J. E., "Image storage and permanence considerations in the long-term preservation of photographic images–update 2010," in *Journal of Physics: Conference Series*, 2010, p. 012008.

[65] Landwehr, N., Hall, M., and Frank, E., "Logistic model trees," *Machine learning,* vol. 59, pp. 161-205, 2005.

[66] Lindley, S. E., Marshall, C. C., Banks, R., Sellen, A., and Regan, T., "Rethinking the web as a personal archive," presented at the Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro, Brazil, 2013.

[67] Lord, P., Macdonald, A., and Committee, J. I. S., *e-Science Curation Report: Data curation for e-Science in the UK: an audit to establish requirements for future curation and provision*: Digital Archiving Consultancy Limited, 2003.

[68] Lynch, C. A., "Institutional repositories: essential infrastructure for scholarship in the digital age," *portal: Libraries and the Academy,* vol. 3, pp. 327-336, 2003.

[69] Manevitz, L. and Yousef, M., "One-class document classification via neural networks," *Neurocomputing,* vol. 70, pp. 1466-1481, 2007.

[70] Marshall, C. C., "How people manage personal information over a lifetime," *Personal information management,* pp. 57-75, 2007.

[71] Marshall, C. C., "Maintaining personal information: Issues associated with long-term storage, preservation, and access," *Personal Information Management: Challenges and Opportunities,* 2006.

[72] Marshall, C. C., "Rethinking personal digital archiving, Part 1: Four challenges from the field," *D-Lib magazine,* vol. 14, p. 2, 2008.

[73] Marshall, C. C., "Rethinking personal digital archiving, part 2: implications for services, applications, and institutions," *D-Lib magazine,* vol. 14, p. 3, 2008.

[74] Marshall, C. C., Bly, S., and Brun-Cottan, F., "The long term fate of our digital belongings: Toward a service model for personal archives," in *Archiving Conference*, 2006, pp. 25-30.

[75] Marshall, C. C., McCown, F., and Nelson, M. L., "Evaluating personal archiving strategies for Internet-based information," in *Archiving Conference*, 2007, pp. 151-156.

[76] Marshall, C. C. and Shipman, F. M., "Saving, reusing, and remixing web video: using attitudes and practices to reveal social norms," presented at the Proceedings of the 22nd international conference on World Wide Web, Rio de Janeiro, Brazil, 2013.

[77] Masanès, J., *Web archiving*: Springer, 2006.

[78] McCown, F., " Lazy Preservation: Reconstructing Websites From," Old Dominion University, 2007.

[79] McCown, F., Marshall, C. C., and Nelson, M. L., "Why Web Sites Are Lost (and How They're Sometimes Found)," *Communications of the ACM,* vol. 52, pp. 141-145, 2009.

[80] McCown, F., Smith, J. A., and Nelson, M. L., "Lazy preservation: Reconstructing websites by crawling the crawlers," in *Proceedings of the 8th annual ACM international workshop on Web information and data management*, 2006, pp. 67-74.

[81] Meneses, L., Barthwal, H., Singh, S., Furuta, R., and Shipman, F., "Restoring Semantically Incomplete Document Collections Using Lexical Signatures," in *International Conference on Theory and Practice of Digital Libraries*, 2013, pp. 321-332.

[82] Meneses, L., Furuta, R., and Shipman, F., "Identifying “Soft 404” error pages: analyzing the lexical signatures of documents in distributed collections," in *Theory and Practice of Digital Libraries*, ed: Springer, 2012, pp. 197-208.

[83] Meneses, L., Jayarathna, S., Furuta, R., and Shipman, F., "Grading Degradation in an Institutionally Managed Repository," in *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2015, pp. 263-264.

[84] Miyata, K., "Issues and Expectations for Digital Archives in Museums of History: A View from a Japanese Museum," in *Archiving Conference*, 2004, pp. 108-111.

[85] Monard, M. C. and Batista, G. E., "Learmng with skewed class distrihutions," *Advances in Logic, Artificial Intelligence, and Robotics: LAPTEC*, vol. 85, p. 173, 2002.

[86] Odom, W., Sellen, A., Harper, R., and Thereska, E., "Lost in translation: understanding the possession of digital things in the cloud," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012, pp. 781-790.

[87] Peters, R., "Exploring the design space for personal information management tools," in *CHI'01 Extended Abstracts on Human Factors in Computing Systems*, 2001, pp. 413-414.

[88] Petrelli, D., Whittaker, S., and Brockmeier, J., "AutoTopography: what can physical mementos tell us about digital memories?," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2008, pp. 53-62.

[89] Phelps, T. A. and Wilensky, R., "Robust hyperlinks and locations," *D-Lib magazine,* vol. 6, pp. 1082-9873, 2000.

[90] Poursardar, F. and Shipman, F., "How Perceptions of Web Resource Boundaries Differ for Institutional and Personal Archives," in *2018 IEEE International Conference on Information Reuse and Integration (IRI)*, 2018, pp. 126-129.

[91] Poursardar, F. and Shipman, F., "On Identifying the Bounds of an Internet Resource," in *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, Chapel Hill, NC, 2016, pp. 305-308.

[92] Poursardar, F. and Shipman, F., "What Is Part of That Resource? User Expectations for Personal Archiving," in *Digital Libraries (JCDL), 2017 ACM/IEEE Joint Conference on*, Toronto, Ontario, Canada, 2017, pp. 229-232.

[93] Rodden, K. and Wood, K. R., "How do people manage their digital photographs?," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2003, pp. 409-416.

[94] SalahEldeen, H. M. and Nelson, M. L., "Losing my revolution: How many resources shared on social media have been lost?," in *International Conference on Theory and Practice of Digital Libraries*, 2012, pp. 125-137.

[95] Seni, G. and Elder, J. F., "Ensemble methods in data mining: improving accuracy through combining predictions," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 2, pp. 1-126, 2010.

[96] Smith, J. A. and Nelson, M. L., "Creating Preservation-Ready Web Resources," *D-Lib magazine*, vol. 14, 2008.

[97] Smith, J. A. and Nelson, M. L., "Generating best-effort preservation metadata for web resources at time of dissemination," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, 2007, pp. 51-52.

[98] Spinellis, D., "The decay and failures of web references," *Communications of the ACM*, vol. 46, pp. 71-77, 2003.

[99] Swan, A. and Brown, S., "Open access self-archiving: An author study," 2005.

[100] Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R., "The perfect search engine is not enough: a study of orienteering behavior in directed search," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2004, pp. 415-422.

[101] Toyoda, M., "The History of Web Archiving," *Proceedings of the IEEE*, vol. 100, pp. 1441-1443, 2012.

[102] Whittaker, S., Bellotti, V., and Gwizdka, J., "Email in personal information management," *Communications of the ACM*, vol. 49, pp. 68-73, 2006.

[103] Whittaker, S., Bergman, O., and Clough, P., "Easy on that trigger dad: a study of long term family photo retrieval," *Personal and Ubiquitous Computing*, vol. 14, pp. 31-43, 2010.

[104] Whittaker, S. and Sidner, C., "Email overload: exploring personal information management of email," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1996, pp. 276-283.

[105] Wilson, A., "A performance model and process for preserving digital records for long-term access," in *Archiving Conference*, 2005, pp. 20-25.

[106] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J., *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2016.