INTELLIGENT INFORMATION INTERACTION FOR MANAGING DISTRIBUTED COLLECTIONS OF WEB DOCUMENTS

A Dissertation

by

PAUL LOGASA BOGEN II

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

DOCTOR OF PHILOSOPHY

December 2011

Major Subject: Computer Science

INTELLIGENT INFORMATION INTERACTION FOR MANAGING DISTRIBUTED COLLECTIONS OF WEB DOCUMENTS

A Dissertation

by

PAUL LOGASA BOGEN II

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

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Richard Furuta
	Frank Shipman
Committee Members,	John Leggett
	Patrick Burkart
Head of Department,	Hank Walker

December 2011

Major Subject: Computer Science

ABSTRACT

Intelligent Information Interaction

for Managing Distributed Collections

of Web Documents. (December 2011)

Paul Logasa Bogen II, B.S., Texas A&M University

Co–Chairs of Advisory Committee: Dr. Richard Furuta
Dr. Frank Shipman

Digital collections are ubiquitous. However, not all digital collections are the same. While most digital collections have limited forms of change – primarily creation and deletion of additional resources – there exists a class of digital collections that undergo additional kinds of change. These collections are made up of resources that are distributed across the Internet and brought together into the collection via hyperlinking. This means the underlying collection members are not controlled by the curator of the collection. Resources can be expected to change as time goes on. To further complicate matters these collections can be hard to maintain when they are large, highly dynamic, or lacking active curation. Part of the difficulty in maintaining these collections is determining if a changed page is still a valid member of the collection. While others have tried to address this problem by measuring change and defining a maximum allowed threshold of change, these methods treat all change as a potential problems and treat web content as a static document despite its intrinsically dynamic nature. Instead, I approach the problem of determining significance of change on the web by embracing it as a normal part of a web document's lifecycle, Instead of using thresholds to identify abnormal changes, I determine the difference between what a maintainer expects a page to do and what it actually does. These

models are created using a variety of feature extractors to find pertinent information in a page, a Kalman filter to model the history of a page and predict a next version and finally classification of results into either expected or unexpected change. I evaluate the different options for extractors and analyzers to determine the best options from my suite of possibilities. This work is informed by a series of studies on both web pages and potential collection maintainers, observations of the NSDL Pathways, and a ground-truth set of blog changes tagged by a human judgment of the kind of change. The results of this work showed a statistically significant improvement over a range of traditional threshold techniques when applied to the collection of tagged blog changes.

To Dorre, E Lee, and Roo

Quae amissa salva

ACKNOWLEDGMENTS

Throughout the course of my dissertation work I have received guidance, assistance, and the occasional smack across the head from a number of colleagues in the Center for the Study of Digital Libraries.

In particular, I would like to thank Luis Francisco-Revilla for his assistance in the initial design brainstorming sessions for the Distributed Collection Manager in 2004 and for his assistance analyzing data as part of my longitudinal study on blogs in 2006. Also, Unmil Karadkar has been a constant sounding board and mentor in developing my ideas and abilities as a young researcher. He has acted as an editor on several of my publications and has been a valuable resource to my work. I'd like to thank Neal Audenaert for enduring my grousing and reminding me to task a step back and pay attention to the bigger picture.

Two people in particular were of great help when I was trying to devise a means to model expectation of change, Dr. Ricardo Guitierez-Osuna and Joshua Johnston. I'd like to thank Dr. Guitierez-Osuna for pointing me in the direction of filter methods and Joshua Johnston for helping implement the prototype that was the basis for our 2008 paper.

Each of these people have been valuable assets and in some cases friends through this long process and I appreciate greatly what each have done.

Finally, I would like to thank the researchers and staff that I have worked with for the Ensemble Project. The project has given valuable insight into real large-scale distributed collections.

TABLE OF CONTENTS

CHAPTER Page

I INTRODUCTION 1

II NSDL PATHWAYS 5

A. AMSER 5
B. Physics and Astronomy Pathway 6
C. Computational Science Education Resources Pathway . . . 6
D. Materials Digital Library Pathway 7
E. A Taxonomy of Web-Based Collections 8
F. Curated Undistributed Collections 9
G. Curated Distributed Collections 10
H. Uncurated Undistributed Collections 11
I. Uncurated Distributed Collections 12

III PROBLEM 15

IV RELATED WORK 18

A. Web Resource Change 18
1. Patterns of Change 22
B. Collections of Web Resources 22
1. Personal Collections 22
2. Subscription-based Technology 23
3. Social Media 24
C. Textual Analysis 25
1. Stylometry 26
D. Genre Theory 27
1. Social Purpose in Genres of Change 29
E. Filter Methods 29
1. Kalman Filters 30

V PRELIMINARY STUDIES 32

A. Patterns of Blog Change 32
B. Personal Distributed Collections Among Social News Users 35
1. Methodology 37

CHAPTER Page

a. Demographics 37
2. Results 38
a. Collection Usage 38
b. Collection Management Techniques 40
c. Subscription Technologies 42
d. Desired Features 42
C. Characteristics of Bookmark Usage Among Graduate Students 43

VI THE DISTRIBUTED COLLECTION MANAGER 46

A. General Architecture 46
B. Nona 47
C. Decima 47
D. Phaeton 48
E. Morta 49
1. Premorta 49
2. OctMorta 50
3. Postmorta 51
F. Hannah 2 52
G. Ananake 54
H. Ianus 55

VII KALMAN FILTER PAGE CHANGE ANALYSIS 57

A. Feature Selection 58
B. Change in Blogs Ground Truth Corpus 59

VIII ANALYSIS 60

A. Feature Selection 60
B. Outlier Detection 62
1. Heuristic Methods 63
2. Classifier Methods 67
3. Statistical Outlier Detection 75
4. Ensemble Method 77
C. Kalman Filter Versus Cosine-Similarity 78

IX CONTRIBUTIONS OF WORK 83

X FUTURE WORK 85

A. Searle's Chinese Room and Expectation Modelling 85

CHAPTER Page

B. Decima 2 86
1. FAR 2 86
C. Phaeton 87
D. Morta 2 87
E. Hannah 3 88

XI CONCLUSIONS 89

REFERENCES 91

APPENDIX A 101

VITA 109

LIST OF TABLES

TABLE Page

I Taxonomy of online collections by how change is managed 9

II Measures used in WebTango (WT) versus the Proportional Algorithm (PA) 21

LIST OF FIGURES

FIGURE Page

1 Absolute Change in Blogs. 33

2 Relative Change in Blogs. 34

3 Change by Day of Week. 35

4 Types of Sites in User Collections. 39

5 Relative Frequency in Which Users Lose Track of Their Collections. . 41

6 Hannah 2: Login Screen. 50

7 Hannah 2: Main Screen. 51

8 Hannah 2: Open Dialog. 52

9 Hannah 2: Cache View Dialog. 53

10 Hannah 2: Feature Extraction Status Dialog. 54

11 Hannah 2: Analysis Scheduling Screen. 55

12 Hannah 2: Analysis Results Dialog. 56

13 Boxplot of Accuracies for Heuristic Methods. 63

14 Boxplot of F-Scores for Heuristic Methods. 64

15 Static Technique Results on Outer Court Blog. 65

16 Constant Step Technique Results on Outer Court Blog. 66

17 Limits Technique Results on Outer Court Blog. 67

18 Boxplot of Accuracies for kNN Methods. 68

19 Boxplot of F-Scores for kNN Methods. 69

FIGURE Page

20 Boxplot of Accuracies SVM Methods. 71

21 Boxplot of F-Scores SVM Methods. 72

22 Boxplot of Accuracies for Discriminant Analysis Methods. 73

23 Boxplot of F-Scores for Discriminant Analysis Methods. 74

24 Boxplot of Accuracies for Statistical Outlier Methods. 76

25 Boxplot of F-Scores for Statistical Outlier Methods. 77

26 Boxplot of Accuracy for Ensemble Method. 78

27 Boxplot of F-Scores for Ensemble Method. 79

28 F-Score and Accuracy at Multiple Relative Threshold Levels for Feature Sets. 80

29 F-Score and Accuracy at Multiple Absolute Threshold Levels for Feature Sets. 81

CHAPTER I

INTRODUCTION

Collections are a part of our daily lives, whether it be the books on our shelves, the files on our computers, or our digital audio collections. Most of these collections have the benefit that the curator, or curating organization, of the collection also possesses the artifacts in the collection. However, there is a class of collections, which I call distributed digital collections, that differs from traditional digital libraries in that the curator of the collection does not possess the artifacts. Without possession of the artifacts, this means that the curator can not control how they change to the same extend that a curator of an undistributed collection can.

By examining the characteristics of different classes of collections, I will build a taxonomy of collections of web resources and the challenges inherent in each class. From this taxonomy I will then focus on three kinds of distributed digital collections: educational resource lists, personal bookmarks, and federated repositories of web resources. This class of distributed collections in general and those named in particular face challenges in maintainance which I believe can be mitigated through modelling a user's expected behavior and identifying deviations in this expected behavior. First, however, I will explore the history of digital collections and the work out of which my work emerged.

Bush's associative trails offer an early glimpse into a form that a digital collection may take [1]. In the early days of the web from 1993 to 1995, NCSA and later Netscape maintained a collection of new web pages [2]. Yahoo! was founded in 1994 to create

---

This dissertation follows the style of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

a hierarchical directory of interesting web sites that, while eclipsed by search engines, is still maintained today [3]. Despite the rise of the search engine, Obendorf et al. found in their study from 2004 to 2005 that revisitation was as likely to originate from a bookmark as from a repeated search [4]. They additionally speculate that in the majority of cases where neither bookmarks nor searching were used, many users were taking advantage of bookmark like behavior from other means, such as, URL auto-completion or retracing previous browsing behavior from a known start state.

The community-centric banner of Web 2.0 has spawned a number of sites such as Wikipedia, Flickr, del.icio.us, reddit and Technorati that seek to build collections, both site-wide and personal, via a community zeitgeist. However, many of these digital collections share a common problem: over time their collections age and the constituents change, become obsolete, or disappear. The result of collection aging has been a diminishing rate of return for the overhead in collection creation. If I can extend the lifespan of these collections, users may be more motivated to create them and thus decrease or prevent instances of rediscovery in the future.

It was in this environment of growing collections that the Distributed Collection Manager (DCM) – the successor to the Walden's Paths Project's PathManager path maintenance utility [5] was conceived. DCM was motivated not only by the observation that the fluidity of web pages leads to collections becoming stale, requiring revisions and updates [6], but also by observations that the web as a whole was changing and that assumptions made by PathManager may no longer be valid [7]. Unlike PathManager, which was focused on maintaining a path in Walden's Paths, DCM is more general and supports other forms of web-based collections, such as bookmark lists and web resource guides. While a path was a well-defined system artifact produced by Walden's Paths, a general web-based collection, as DCM envisions, is a poorly-defined social artifact.

To help inform the design of DCM, I conducted two studies. The first study recruited users of social news sites and solicited them about their habits regarding their personal collections including tools they use, kinds of sites they visit, and issues they face in managing their collections. I followed this study with an examination of personal collections themselves, primarily bookmarks. My studies affirmed the existence of personal collections of highly dynamic web sites, the inadequacies of existing tools to maintain these collections, and the importance of content (both textual and visual) over presentation and interactive features of a site.

Additionally, since 2006 I have been building a test set of caches of a wide variety of popular blogs collected four times a day. From this collection, I used human judges to evaluate a selection of blogs that potentially contained abnormal changes. This tagged test set is then used as a ground truth to evaluate features and techniques to measure change.

From this work I have found that the application of the Kalman Filter provides a more flexible means of modelling expectation of change and identifying abnormal changes than traditionally threshold-based change metrics. In particular, my method can detect changes that threshold-based change metrics can not, such as, abnormal lulls and bursts of activity in addition to the large amounts of abnormal change that threshold techniques detect.

I have also begun to work with community-created collections through the Ensemble Project – a multi-university project funded by the NSF to add a computing education oriented portal to the NSDL family of STEM Pathways websites.

One aspect of this effort is the creation of tools to support these focuses. Since DCM is a tool to support the maintenance of collections, it provides the Ensemble project a tool to maintain personal collections of computing resources.

Another aspect of Ensemble is the creation of collections of web-based materials

to support the areas of focus. These collections are distributed not only in terms of the members being distributed across the web, but the collections themselves are spread out across the institutions collaborating on the project. The widely distributed nature of these collections makes maintenance very difficult. In fact, what a collection contains may be ambiguous as some sub-collections may be maintained by communities that are not directly involved with the Ensemble project. In response to this, DCM is being adapted to help maintain these highly-distributed collections.

A third aspect of Ensemble is the incorporation of a number of non-traditional resources including social networking sites and computing media. This combination of traditional and non-traditional elements yields a new model of understanding of social media.

The remainder of my dissertation will begin with a discussion of the different kinds of collections contained inside NSDL Pathways. This will be followed by the presentation of my taxonomy of collections. From there I will explore related work, followed by the issues of distributed collection management I have identified. This will lead to descriptions of the design of my two studies. From these studies I focus on one genre of highly dynamic web pages, blogs, which I have successfully analyzed and produced models of expectation that are statistically better at finding abnormal change. Finally, I will draw conclusions and identify my next steps.

CHAPTER II

NSDL PATHWAYS

The Ensemble Project has partnered with a number of existing computing resource projects. These projects each maintain collections of varying characteristics [8]. The Algorithm Visualization Portal (Algoviz) contains a rich collection of distributed resources with local metadata [9]. Submission of these items is done by registered members of the user-community that has vested interests in the portal and each user curates their resources. For instance, the Computing and Information Technology Interactive Digital Educational Library (CITIDEL) is a centralized collection of documents drawn from snapshots of distributed sources [10]. Submissions to CITIDEL are recommended by the user-community but approved and edited by a set of maintainers [10]. However, maintenance of these entries is still a responsibility of submitters. The Computer Science Syllabus collection is another example of a centralized collection of documents [11]. However, the Syllabus collection is a collection of static examples of current and past syllabi and currently is curated by the maintainers [11]. Beyond the projects inside the Ensemble Pathway, other NSDL Pathways have collections of interest to DCM.

A. AMSER

AMSER, the Applied Math and Science Pathway, provides references to tens of thousands of Web-based resources covering the Library of Congress Classification taxonomy and extends beyond the Pathway's focus of applied math and science into areas like religion, art, and philosophy [12]. Each resource has a community metadata curator who was the original submitter of the resource to AMSER, a third-party curator who hosts the content, and a author who created the content in the first place. This

means that for each resource the owner may not be affiliated with AMSER and for decisions made at the level of the project would be unacceptable. These references are produced by a range of providers, including large corporations (e.g., the BBC), university departments, and individual teachers. In the top-level category of art, there exist 155 resources ranging from specific lesson plans, government documents, research pages, companies providing educational services and organizations.

B. Physics and Astronomy Pathway

ComPADRE, the Physics and Astronomy Pathway contains thousands of resources in multiple hierarchically arranged repositories containing metadata and links to external resources submitted and maintained by students, teachers, researchers, and the general public [13]. This wide distribution of curation responsibility increases the potential that the resource may suffer from benign neglect as a submitter may not accept the assumed responsibility that submission entails.

C. Computational Science Education Resources Pathway

CSERD, the Computational Science Education Resources Pathway contains thousands of references maintained in overlapping communities [14]. Submission is open (with administrator review) to all. Maintenance of resources is a process where a visitor notices an issue, provides corrected metadata and submits the correction to the pathway administrators who then take corrective action. The result of this process is that there is no well-defined curator and resource validity is left up to information seekers who find a resource. The result of this is a resource that is no longer valid is not discovered until an information seeker fails to find the resource they were looking for. In this situation the pathway administrators are left to hope that an informa-

tion seeker opts to inform them of the issue and possibly provide a resolution to the problem instead of moving along to another resource or even to another site.

D. Materials Digital Library Pathway

MatDL, the Materials Digital Library Pathway is divided into eight communities that control membership to their communities [15]. Community members can then submit resources to a number of repositories controlled by each community. Each resource has a set of locally stored attached files and a set of external URLs that provide related information to the resource. By making external resources a second-class member of the collection that only support the first-class stored documents, the communities may have less incentive to maintain these external links as they may not want to dedicate as much time to these second-class members as they do to the first-class members.

E. A Taxonomy of Web-Based Collections

From the various collections that make up the NSDL Pathways, I have observed basic characteristics of collections that inform how these collections change over time. These characteristics can be used to define a taxonomy of primal kinds of web resource collections in regards to change.

When examining the collections that make up the NSDL Pathway projects I noted that each collection took a different approach to how materials are identified and incorporated into the collection and how maintenance is performed on the collections.

Some collections such as AMSER depend on community submissions, while others such as the MatDL have submissions restricted only to administrator vetted contributors. Likewise how the pathways incorporate submissions vary from distributed collections, like AMSER and the CSERD, to undistributed collections like, the MatDL and CITIDEL. Additionally, some pathways such as Ensemble and ComPADRE are instead federated collections where the constituent members could be distributed or not.

In some systems, such as Algoviz and AMSER, curatorship and authorship are done by the same person. In these systems there is a measure of reliability of the author-curator performing her role. However, in other systems such as CITIDEL and ComPADRE, the curatorship is left to the wide user community who may not share the sense of personal responsibility an Algoviz or AMSER author assume. Finally, a few, such as CSERD, opt for a highly controlled curation process where the site administrators perform all curative actions.

From these observations I have sectioned the space of possible collections into four different groups based on whether they are curated or uncurated and if they are distributed or not. This creates a taxonomy of four types of collections, as far as

Table I. Taxonomy of online collections by how change is managed

	Curated	Uncurated
Distributed	Expert evaluation, changes beyond control of maintainer. (Web resource lists, social bookmarking)	Evaluation done by visitors, changes beyond control of evaluators. (External links on Wikipedia, social news sites)
Undistributed	Expert evaluator, changes controlled by creator. (Traditional corporate web-based collections)	Evaluation done by visitors, changes controlled by creator. (Traditional personal web-based collections)

management of collections is concerned. In addition to the discussion in the remainder of the chapter, a summary of the taxonomy can be seen in Table I.

F. Curated Undistributed Collections

The first kind of collection type is a curated undistributed collection. This is the simplest kind of collection to manage. The maintenance of resources is well-defined and the contents of the collections are consolidated with the collection. This kind of collection does not suffer from link rot or other kinds of change events. Change in the collection is limited to resources being added, deleted, and purposefully modified by an expert with the responsibility for curation. This is the traditional digital collection that is common to digital libraries of all kinds. This kind of collection also includes a majority of non-digital collections. Non-digital forms include most museums, art galleries, and archives. To understand these collections I will revisit an example from the NSDL Pathways, the MatDL.

As mentioned previously, the MatDL is a closed set of eight communities each

curating their own collections. Each entry in their collection has clearly assigned authors, a submitter, and any modifiers. The submitter is the first level curator and often the author, thus providing incentive to maintain the entry. In additional the broader community acts as a wider second level of curation. Membership in a group signifies a commitment to the broader curative responsibilities and submission refines that to an explicit commitment. Despite these strong commitments by the community members and submitters, the task of curation is of low impact since it is primarily revision to mistakes in metadata or replacement of links to the second-class external resource pointers.

G. Curated Distributed Collections

Next I have a collection that may seem similar to the first in many ways, except the collection is distributed. These collections are not as common as a curated undistributed collection and again have a non-digital analogue. Libraries traditionally were undistributed but due to interlibrary loan agreements, online database access, and the multi-branch systems that are now commonplace they are in many ways more a distributed collection than not. Some larger Museums, such as the Smithsonian, and the J. Paul Getty Trust and large archives, for example, the National Archives, also represent examples of curated distributed collections. Algoviz is an example of a digital curated distributed collection.

Algoviz stands in contrast to the MatDL in that only metadata is stored internally to the collection. In one sense, the collection is only a set of metadata records. However, due to the ease of linking on the Web, one of the metadata fields is the resource's URL thus allowing easy access to the material. While this entails benefits for the collection, such as not incurring the cost of hosting and maintaining the resources,

the collection must sacrifice control over the resources and depend on an externality that may not share the same desires, purposes, or notions of importance that the curator does. Thus the curator needs to be aware of the status of the members of her collection and be proactively warned when a potential problem arises so she can take action to preserve or replace her resource before it is too late.

Fortunately in this class of collection, there is still a well-defined curator whose expert domain knowledge can be leveraged to help with tasks that may be computationally complex or impossible.

H. Uncurated Undistributed Collections

An uncurated undistributed collection is at first glance an odd idea. However, these collections are arguably the most common and often unrecognized as collections. For non-digital examples the key is that these collection are lacking in either selection, caretaking, or presentation. Hobbyist collections sometimes fall under this category as often they are built more by visceral attraction than a purposeful decision for inclusion. Once the collection is assembled, they are often stored in closets, shoeboxes, or dusty shelves. Finally some collections are never organized and are mistaken by others as junk or clutter. Personal collections of photographs are a prime example for this kind of benign neglect. How many of us have shoeboxes full of photographs, many of them blurry and effectively useless? For digital collections these are typified by poorly organized directories of music, pictures, or documents. Online these collections are common on personal web sites and Web 2.0 sites that allow open submission of materials. In the NSDL Pathways there is not a clear example of this kind of site.

However, some of the curated undistributed collections with looser curation standards, such as CITIDEL and ComPADRE, have the potential to behave as if they

were uncurated since curation responsibilities often fall on submitters who may not have vested interests in maintaining their submissions. Thus, these sites have a greater likelihood of inaccurate and insufficient metadata than their more actively curated cousins. These sites also have a potential of growth that may make a future attempt to curate and bring the site under control a time consuming tasks. One option, and the traditional approach to bring loosely curated and uncurated collections into a state of stricter curation, is to create tools to help automate or streamline the hard task of *ex post facto* curation. However, tools to help submitters fulfill their implied commitment to curation may prove to be essential in preventing these collections from experiencing the drawbacks of an uncurated undistributed collection in the first place.

I. Uncurated Distributed Collections

Finally I get to the strangest form of collection. These collection are lacking in curation and do not have a single organizing location. One way of thinking of these collections in the non-digital world is to see them as a potential collection. Before the information age, these collections were purely thought experiments. While many would say that unrealized collections are not collections, you could argue that since the characteristics that make a collection cohesive are intrinsic to the collection. Thus, the collection essentially always exists.

In philosophy there exists a thought experiment called the Ship of Theseus [16]. According to legend, the Ancient Athenians were have said to have continued to maintain Theseus' ship for centuries after his supposed life as proof of his existence. Over time every original component of the ship had been replaced. This leads to the question that if eventually no original piece of the ship remained, was it still the ship

of Theseus? The Aristotlian solution to this dilemma was dependent on examining the causes that defined the object [17]. According to Aristotle there are four causes: the formal cause – or the design, the material cause – the physical components, the final cause – the purpose, and the efficient cause – the process of creation. For the ship of Theseus two of these, the formal and the final cause, still hold and the efficient cause may hold if care was taken to use the same methods to make the repairs. Since two (and maybe three) of these causes still hold you could conclude that, yes, the ship was still Theseus' ship.

Likewise a potential collection is a collection that has not been realized as the material cause (the collection members) and the final cause (the purpose of the unrealized collection) hold even before the collection is created. The efficient cause may also hold as it may be possible to create a collection earlier than when it is eventually created. Much like the Ship of Theseus, a potential collection's existence as the collection is only called into doubt due to the lacking of one cause, unlike the material cause of the ship, here the formal cause is lacking as a potential collection is by definition undesigned since it has not undergone the process of selection.

Once we move to the digital realm, uncurated distributed collections become more obvious. Social news sites are a growing example of these kinds of collections. On social news sites, such as reddit, user submissions drive the growth of their collection of online resources. Once submitted, other than spam or rule violations, the submissions are published without review and cannot be modified later. Voting on resources may propel the link, and attached commentary, to a short-lived prominence on the site, but as time goes on the resource descends into obscurity. Nowhere in the process is a resource's continued existence verified and often links are very transient, such as a humorous mistake on a Wikipedia entry. The only curation activity that these kinds of collections undergo is the activity of assembling personal lists of saved

resources that members create, but cannot share.

For my work, I am focusing on distributed collections, both curated and uncurated. For curated collections I intend to help ease the burden of curation while for the uncurated ones, I intend to encourage their transformation into curated collections. The next chapter will deal with the particulars of the problem and the way in which I approached the problem.

Currently, the NSDL does employ a tool, known as the Vitality Checker, to help maintain the distributed resources in the Pathways [1]. However, this tool uses a basic method that only checks if a 200 status code is returned on connection to a URL. Other systems that have been designed to help manage collections of web resources are discussed in the Related Work chapter.

[1]Information on the NSDL Vitality checker was obtained via source code provided courtesy of the UCAR Digital Learning Sciences group.

CHAPTER III

PROBLEM

> It is the invariable lesson to humanity that distance in time, and in space as well, lends focus. It is not recorded, incidentally, that the lesson has ever been permanently learned.
>
> Isaac Asimov, *Foundation*

While the post-creation activities around a distributed collection differ depending on whether they are curated or not, the process of creation is often very similar and sets the stage for both the previously stated goals of easing management and encouraging collectors to opt for a curative process.

In both cases, initially a collector identifies a number of sites they are interested in defining a collection around. This collection has an intended purpose that effects what a collector's expectations are from the sites and what level of commitment the collector would be willing to make to the maintenance and curation of the collection. Due to this, a list of members of a collection is insufficient for management. For instance, a user who creates a collection on bats that points to www.cnn.com is more likely interested in an article on the site at the time of collection creation, while a user who creates a collection of on-line information outlets would have www.cnn.com as an instance of an information outlet. Both of these cases differ from the case where a user who includes www.cnn.com to help them stay informed of news events. When www.cnn.com changes, the bat collection may no longer be coherent as an article on any number of topics could now be on www.cnn.com. However, the second collection is still coherent as long as www.cnn.com is a news site. Only a change in genre or the page going static would be considered abnormal. Finally the third collection seeks out major news events would desire notification of behavior that indicates breaking

news events. Examples of this case include during 9/11 when CNN's normal constant flow of articles halted while they were trying to sort out what was going on, or after Michael Jackson's death when entertainment news sites were rapidly posting every bit of information they received out of a concern that they would be perceived as disconnected from an unfolding event.

Thus, my problem is defined as trying to identify changes that may be considered as abnormal in terms of what a user would expect when revisiting the page. Note that this is essentially the visceral reaction one would have when re-checking a site as part of a curator's maintenance task. The goal is to allow the first step of maintenance re-visitation to be done automatically either at a frequency more often than or to more resources than a curator normally would have the time to maintain.

Determining this context is a hard task itself and beyond the scope of this work. Therefore I require a user to indicate a purpose for a resource to be analyzed. Once this is set the system begins collecting a set of samples of the page over an initial time-frame. Using these initial samples it tries to model the behaviors of change for each page. At this point the system continues to periodically reevaluate the page. These evaluations utilize a combination of metrics on the content, context, and structure of the page over time.

When a change is determined to be out of the expectation, the system then notifies the user who can then either agree or disagree with the evaluation. This feedback is used to adapt the analysis for the collection. Likewise, if a change is missed, the user can browse the history of a page and indicate to the system an undesired change that had been treated as expected.

To address the problem of management of distributed collections of web-resources, we have designed a system that can both assist collectors in opting to curate their collection and help ease the burden on curators of digital collections.

Before describing the details of our design, I will discuss works related to my work and describe four preliminary studies that have informed my work.

CHAPTER IV

RELATED WORK

My work draws on several areas of study both in Computer Science and in the Humanities disciplines of Textual Analysis and Genre Theory. This chapter will begin with a discussion of other work in the area of web resource change followed by discussions of work focused on social media, subscription-based technologies, and personal collections. This will be followed by a discussion of the textual analysis and genre theory work that helped shaped my approach. Finally, a quick background into the primary analysis technique and its class of pattern recognition algorithms will be given.

A. Web Resource Change

Collection management focuses on the problem of managing change in a collection. Management of collections has been an important topic in many research communities even before the advent of information systems. In the past, change has been viewed from a perspective similar to that of a traditional library – new documents arrive and old documents are lost or are worn out, but the existing documents remain unchanged [18]. This perspective of change only contemplates types of change whose magnitude and frequency are very low since changes in the document contents are not possible except as a consequence of wear, damage, or annotation. As a result, the manual maintenance of the collection is possible in traditional libraries, even large ones.

This perspective of change has carried over to digital libraries. This is a reasonable approach not only in the traditional library, but in curated digital libraries, since they both share a fairly tight control over their collections; sometimes even more so in the digital library where wear is not as great of an issue. However, when the contents

of the members of a collection are loosely controlled, like a collection of web resources, this approach does not apply. The source of this incompatibility is the fact that many new types of change not only can occur, but often do. These changes are not only in membership to a category like in a curated collection, but can also be in what the individual documents contain or how they are presented. Additionally, the frequency of change can increase to unmanageable levels. Initially, the majority of responses from the community were to find ways to "deal with change" [18, 19, 20, 21, 22, 23]. These solutions essentially try to resist change. While more recently, the community has grown more aware that change is intrinsic to digital collections, it still seeks to only deal with change, as illustrated by existing work such as Askehave and Nielsen's work on digital genre where they saw change as a detrimental factor that could render their determination of characteristics of a genre invalid as time went on [24]. An early work to try and examine change on the web was the Do-I-Care Agent (DICA) [18]. DICA was primarily interested in finding new information for particular users. Changed documents were treated as if a new document had been entered in the user's collection. Later, Ashman described three classes of strategies for dealing with change on the web: preventative, corrective, and adaptive [19]. Preventative strategies are applied before-hand to avoid change from the beginning, such as caching. This makes the document static. Corrective strategies are applied after a problem and are designed to find replacements or equivalents. Adaptive methods are those that try to hedge the maintainer's bet by taking preventative actions that can then be used to form a corrective strategies. These classes still assume that change is a problem to be resolved and not the nature of the document. In 2001, the Walden's Paths project managed change by extending Johnson's approach for distance between page versions to determine how much a page had changed [20]. Further examples, can be seen in Koehler's change study [21], Boese and Howe's study that viewed change as a possi-

ble detriment to classification work [22], and Ivory and Megraw's study of Web Site Design Patterns [23]. In each of these cases the authors tried to factor out change in their collections. Their goal was either to find a underlying truth about the document without change getting in the way, or to restore the integrity of a collection *tainted by change*. While Askehave and Nielsen recognized change as an integral part of web documents, they admittedly did not fully account for their "view of web-mediated genres as dynamic documents" [24]. In fact, no change management systems have been built around the changing nature of web documents.

Despite past tendencies of treating change as an aberration to correct, the previous work does provide good sources of metrics for measuring different types of change. In particular Ivory's dissertation on WebTango provides a large set of metrics to measure features of pages [25]. While a large number of the metrics are better suited to her task of finding "good web pages" than to building a system around change, several of them do apply the the task of change management. Of these measures fifty-three of them only have meaning in the scope of an entire web site and not an individual web page. Sixty-one are purely presentational and provide no insight into the content of the page. Six of them are measures involving images and their quality which are beyond the scope of my work. However, the remaining thirty-seven measures are for measuring aspects of the text and links in a web page. Several of these measures do not provide additional insight into the content of the document, leaving us with four link measures and eight text measures. These measures are enumerated in Table II.

The Walden's Paths project has created both the Proportional Algorithm for analyzing structural changes in a page [6], and an algorithm for determining page change in context of the rest of the collection [26]. Table II lists out the applicable measures that Ivory identified in her dissertation [25]. Some of these metrics were not used in her WebTango system. While overlaps do exist between Ivory's and

the Walden's Paths' approaches, each individual set of metrics is not comprehensive. Currently, the Proportional Algorithm performs analysis on the Link Count, Word Count and Display Word Count; but does not offer the same coverage that Ivory's eleven applicable measures do. Thus by modifying and combining these approaches I can construct a more comprehensive approach to measuring change.

Table II. Measures used in WebTango (WT) versus the Proportional Algorithm (PA)

Measure	Description	WT	PA
Links	All links.	✓	✓
Page Links	Links to other sections within the page.	✓	
Internal Links	Links that point to pages within the site.	✓	
Redundant Links	Links that point to the same page as other links.	✓	
Words	Visible words.	✓	✓
Overall Page Title Words	Words in the page's title.	✓	
Body Words	Words that are body text but not headings or links.	✓	✓
Display Words	Headings that are not links.	✓	✓
Display Link Words	Words that are both link text and headings.	✓	
Link Words	Link text that are not headings.	✓	✓
Ad Words	Words indicating ads.	✓	
Paragraphs	Subdivisions of the body.		✓

1. Patterns of Change

When change is no longer in opposition to the management philosophy of a digital collection, but an essential part of the collection, the opportunity arises to abstract changes into the broader form. Results from the previously cited methods provide a two-point comparison, which in past work usually [6, 22, 26] was the extent of the comparison. However, when these comparisons continue over a large period of time, patterns of magnitudes and frequencies of change emerge. These patterns provide a strong basis for identifying documents that may not be coherent with their original collection. After conducting a study of blogs from late 2006 to 2007 [7], I concluded that evidence of a commonality of change behavior across a given fine-grained genre, particularly blogs, indicates that associating a pattern of change to a genre enables pages to be grouped by observed change. This better helps to determine an expectation of change. Despite the changing nature of web pages, existing genre-based approaches have treated them as a new class of *static* documents [22].

B. Collections of Web Resources

1. Personal Collections

Since Vannevar Bush's *As We May Think* introduced the concept of a electronic bookmark as a coded index into a microfilm book stored inside the Memex [1], the concept of a bookmark has been an important component of digital collections and hypertexts. This line of work has continued into the present with a wealth of work studying bookmarks.

Li *et al.* [27] were able to point to prior work showing that users did have difficulty keeping things found and organizing information. However, they did not address how people were trying to organize information and if bookmarks were even being used.

Kellar *et al.* 's [28] study into how people seek information on the Web gathered their data by collecting bookmark files and was thus unable to give in insight into what all users were doing as opposed to what users who used bookmarks were doing. The prevalence of bookmarks has been examined three times. First, a 1998 study found that 98% of attendees at an academic conference focused on the Internet had bookmark collections [29]. Jones *et al.* [30], in their 2001 study on how users kept previously found items on the web found showed that only one of their eleven participants used bookmarks. However, their work was focused on single tasks which in many cases were poor examples for normal behavior, such as a collaboratively authored paper and a person using a shared-access machine they didn't use for their regular work. Lastly, a study in 2005 on members of ACM's SIGCHI mailing lists found that 92.4% of the participants created bookmarks [31]. However, other work that examined actual usage of bookmarks through click tracking [4] concluded that people don't revisit bookmarks very often. This seemingly contradictory situation has not been addressed. Why do people create bookmarks, if they are not using them?

2. Subscription-based Technology

Subscription Technologies, such as RSS and ATOM, are technologies that allow a simplified content and metadata feed to be harvested by a system for reuse in another context. These feeds are typically dynamically generated so that a retrieval of the feed always produces the latest content. Subscription technologies continue to be a large area of ongoing research in Computer Science, including previous work by Walden's Paths investigating RSS as a means to automatically augment existing paths with relevant information [32]. Despite their popularity, Liu *et al.* [33] found that while there was a large body of work using RSS as a resource or a tool, there was little to no work about how readers of RSS feeds were using them. In response to this, Liu

delved in to topics such as how many feeds readers read and how frequently their aggregation utilities retrieved the feeds.

3. Social Media

With the rise of the Social Web there came a new approach to bookmarking and news gathering on the web. Social bookmarking and social news sites bring what were once individual activities by a sole user, in the case of social bookmarks, or an editor, in the case of social news, and instead allow a community to identify interesting and relevant resources for each other. Often this involves community voting or tagging to build these rankings. In the realm of social bookmarks, a large portion of the existing work has focused on how the social bookmarking sites can be used to inform other tasks. These range from using social bookmarks to build summaries of web sites [34] to semantic web research attempting to generate ontologies from tags that users had applied to their bookmarks [35]. Another major set of social bookmarking work focuses on the social aspects. Work in this area has delved into topics like the quality of tags [36] and how social networks evolve [37].

Of particular interest to my work, is prior work that attempted to answer the questions "Why do people create tags?" and "What do people use social bookmarking cites for?" The first question was addressed by Lee's work examining motivations for tagging on del.icio.us [38]. In this work a relationship between a person's tagging activity on del.icio.us and the size of their friend list on del.icio.us was found that suggested that people with larger friend lists were more likely to be an active tagger.

Much like the related social bookmarking sites, social news sites, like reddit, digg and fark, consist of user-found links shared amongst a community. Unlike the social bookmarking sites, social news sites have an emphasis on current events and new content. Work on social news sites have been particularly focused on the social

aspect of the sites. For instance, Lerman *et al.* analyzed voting patterns on digg [39].

Throughout the body of work on social news and social bookmarking three questions are not being asked. Are the collections that users are generating important to them? Are they managing their collections? And, do social news and bookmarking sites compliment or supplement bookmark files and subscription technologies?

C. Textual Analysis

In *Cybertext*, Aarseth attempts to show the range beyond traditional literature and hyperfiction that literature can take [40, 17-23]. In order to accomplish this, he defines a number of variables over which literature can range and performs a survey over works that he considers cybertexts. His goal is to illustrate how much wider the field of ergodic literature is than traditional literature [40]. To classify texts, he defines eight variables: dynamics, determinability, transiency, perspective, access, linking, and user functions [40]. Ten of the twenty-three works he surveys are non-digital and provide almost the complete range of possible variables except for transience and texton manipulation. These works include experimental prose and poetry, gamebooks and even *Moby Dick* [40]. Under his definitions, the realm of "traditional literature" falls into a very narrow range in his spectrum: static, determinable, intransient, controlled works with no linking and only an interpretative function. Alternatively, Crystal devises a spectrum from speech-acts to text-acts where spoken language lies on one end and Aarseth's non-ergodic traditional literature lies on the other [41]. To Crystal, the web is neither fixed nor transient, textual nor spoken. Aarseth and Crystal both conclude that web pages are different kinds of texts than what traditional textual analysis deals with [41]. For Aarseth its potential non-linearity and indeter-

minability opens new possibilities for analysis and action [40]. Likewise, Crystal sees web documents as existing in a place between speech and text. This is advantageous since techniques used to analyze both can be utilized. According to Crystal, Internet documents resemble fixated oral performances in many ways [41]. Again this provides opportunity as fixated oral products are the way scholars already work with oral traditions [42].

Many techniques have been used previously to measure change of web documents. Some projects, such as the AT&T Internet Difference Engine [43], have relied on presentation of differences using a traditional differencing algorithms. Others, such as Zoetrope, focus on presenting a user with changes to specific directed portions of the page [44]. Additionally, Greenberg and Boyle used image comparison techniques to identify visual changes between versions of web-based documents [45].

Some have attempted to compile comprehensive lists of change metrics. Ivory and Megraw identified over 150 metrics ranging from traditional text metrics to information about styling, graphics, performance, and linkages [23]. Yadav *et al.* [46] identified four categories of changes: content/semantic, presentation/cosmetic; structural; and, behavioral.

1. Stylometry

Stylometry is a technique used in textual analysis to try and profile the author of a passage. This profile can then be used to support or oppose the notion that two passages were authored by the same person [47]. Stylometry takes an ensemble approach to features; everything from term vector spaces to n-gram analysis, from part of speech comparisons to heuristics based on certain kinds of words and the characteristics of words used. Stylometrics has been used on web sites in the past. Lex *et al.* evaluated different stylometric features to classify genres of blogs and concluded that

while term vector techniques provided better results in most cases, they were more sensitive to topic shifts while their stylometric techniques were less sensitive to topic shifting [48]. Features used by Lex *et al.* included punctuation counts, emoticon counts, distribution of sentence lengths, average words per sentence, number of minimal length sentences, number of unique part of speech tags, and several others. Of their metrics, the best performance was obtained from percentage of tokens identified as adverbs. One factor lacking in Lex's work is they do not examine if a combination of features would yield better results.

D. Genre Theory

Historically the study of genre has been a narrow field of primary interest to the study of literature and pedagogy [49]. However, Australian and American studies in the early 1980s recognized that genres existed beyond literature and could allow instruction of writing skills beyond imitations of others [50]. The same kind of realization that extended the concept of genre to all written communication was repeated in the early 1990s to consider new media [50].

By removing the constraints of the paper medium of books, a more generalized framework was created. This framework has now been specialized to previously unconsidered media. Now, in the 2000s, people reapplied genre to web pages. However, the essential nature of this medium, change, was not being utilized and instead web pages are treated as a new class of *static* documents [22].

If a pattern of change meets the definition of a genre of organizational communication, a three level genre model, as described by Askehave and Nielsen, can be used to identify the core features of the genre [24]. The three levels are the communicative purpose, the move structure, and the rhetorical strategies. Communicative purpose

focuses on the social purpose that a genre addresses. The move structure is the set of functional units common to members of the genre. Lastly, the rhetorical strategies, are devices used in the actual content that are common to a genre.

When I adapt Genre Analysis to the study of web document change, portions of the patterns would have to satisify each of the three levels. Thus, an analysis based on change would have to analyze the shifts in social purpose, the change in the structure of the page, and an analysis of the change of the text itself. Using these aspects on an instance of a document provides an instantaneous genre. The instantaneous genre is the genre of the document as it was at the time it was captured. For example, a blog that normally covers an individual's opinions about technology could actually be giving first hand accounts of a current news event. At that moment the site is acting as a current events news site. While across all instances the site is mainly a blog on technology. The set of all genres at each instance therefore forms a change genre.

The move structure and the rhetorical strategies are well-known domains of textual analysis addressed by discourse analysis, part-of-speech tagging, and term frequency analysis [51]. However, determining social purpose is a much more difficult task that many suggest is not even possible. Since genres are a hierarchical structure, it seems to follow that for every specific genre there is a metagenre such that all members of the metagenre have the same move structure and rhetorical strategies, but different communicative purposes. This metagenre is the implied target of most computational applications of genres studies where the goal is to trim a set of documents to a easier to manage set to hand-sort for a specific purpose (spam filtering, homeland security).

In Krippendorff's seminal work on textual analysis [52], he names six essential questions that must be addressed when analyzing content: Which data is being analyzed? How is the data defined? What population does the data come from? What

is the context of the data? How is our analysis bounded? What are the inferences targeting? By attempting to address these questions in the mindset of Aarseth and Crystal, I hope to provide a sound analysis of web resources.

1. Social Purpose in Genres of Change

As linguistics became more interested in genre during the 1970s and 1980s, one applied linguist, John Swales recognized a failure in not only emerging new genre studies as mentioned previously, but in the traditional genre studies of prose and poetry. In his seminal work, *Genre Analysis: English in academic and research settings* [51], he extends the individual case examples of other field studies that suggested that in selected fields a genre of a document was determined more by social context than by "inner-directed cognitive processes". In fact, Swales, suggests a course of action, namely, treating text as a socially situated-action through speech act theory. Recently, Goldstein and Sabin, used genre identification to classify emails[53]. When reviewing various formulations of speech act theory and shallow discourse annotation systems to facilitate speech act analysis, they identified a comprehensive list of social purposes for statements. By selecting particular speech acts that they felt were of great significance to email communication and developing a series of broad heuristics, they were able to build a rule system for approximating the social purpose and thus the genre of a given email.

E. Filter Methods

According to Jazwinski [54], filter methods can be dated back to work by Gauss on modelling the orbits of the planets. Gauss however was interested in building a deterministic system that minimizes error. Modern filter methods instead take a

probabilistic view that produces a likely range of values. These very in complexity. Kalman Filters are generally seen as the simplest non-stationary filter method to be of practical use; there exists a simpler method by Swerling but it assumes no noise. Contemporaneous to Kalman's linear non-stationary filter with Gaussian noise is Stratonovich's non-linear non-stationary filter with Gaussian noise. Since these filters were developed in the late 1950s and early 1960s, more work has been done creating filters of greater complexity and broader applications than the early methods. For my work I am focusing on the Kalman filter itself as it is a good guideline to start with a simpler model and add complexity only when it is warranted.

1. Kalman Filters[1]

Kalman filters are used to model systems that change over time, allowing predictions to be made about future states [55]. The underlying assumption in these dynamic systems is that future values depend on past values in a linearly. Observations, which may be noisy, can be made on the system, with some information remaining hidden (imperfect information). The state of the system is denoted with the variable x, which may have many hidden states. The state at a given time step t is denoted x_t. Observations about the system are denoted with the variable y_t. x_t and y_t are controlled by the following system of equations:

$$x_t = Ax_{t-1} + v_t, v_t \ \acute{I}(0, Q) \tag{4.1}$$

$$y_t = Cx_t + w_t, w_t \ N(0, R) \tag{4.2}$$

A is a mixing matrix that determines how the system changes from time step $t-1$

[1]©ACM, 2008. This section is a minor revision of the work published in the Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital libraries (2008) http://doi.acm.org/10.1145/1378889.1378941

to t. v is Gaussian noise introduced into the system at each time step, distributed with a mean at 0 and a covariance matrix Q. C is a mixing matrix that determines how we are allowed to view the observables in the system. Observations also include Gaussian noise, w, distributed with mean at 0 and covariance R. The canonical example for describing the use of a Kalman filter is in the tracking of an object moving through space. The state vector x might contain the current position of the object, as well as the velocities along each dimension. In 2D, x might look like $[x_1, x_2, dx_1, dx_2]^T$. The mixing matrix A would be formed such that at each time step, the position in each dimension is updated by the velocity. Say we are able to observe the position of the object in a 2D space (with noisy measurements), but not the current velocity of the object. Then, the matrix C would filter out the velocity values and give us a vector y containing $[x_1, x_2]^T$.

Recently, Kalman filters have begun to be used outside of their traditional application to sensor data modeling and are being applied to textual analyses. In particular, the area of topic tracking has seen great success using Kalman filters. Simultaneously, work has been conducted by Krause, *et al.* , on topic intensity [56], Wang and McCallum on topic trends [57], Cselle, *et al.* , on topic tracking in emails [58], and Blei and Lafferty on topic tracking in Science [59]. Additionally, Van Durme, *et al.* , have used Kalman filters for semantic parsing for question answering systems [60].

CHAPTER V

PRELIMINARY STUDIES

In order to inform the design of my system, I conducted three preliminary studies. The first explored the patterns of change in popular blogs, the second was a survey of users of social news sites, and the third was an examination of bookmark collections belonging to graduate students at Texas A&M University. Each of the studies has helped shaped my work and the design of my system.

A. Patterns of Blog Change[1]

Since documents that are expected to stay static are not as difficult to manage as changing documents, I decided to focus on a genre of web pages where frequent change is expected, blogs. Starting with the 100 most popular (as defined by incoming links) blogs listed by technorati.com in 2006, I discarded all non-English and "dead" blogs to focus on my goal of designing models of change patterns based upon this study. I focused my work on 62 of the 100 blogs[2] from September 25 to December 10, 2006.

Each blog was analyzed with a weighted term frequency based vector-space model. My algorithm measures the degree of change between two versions of a Web resource. This change is measured in terms of degrees. 90 degrees indicates a com-

[1]©ACM, 2007. This section is a minor revision of the work published in the Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital libraries (2007) http://doi.acm.org/10.1145/1255175.1255201

[2]Of the 23 foreign language blogs that I encountered, 20 were not in a Latin script, such as Chinese and Japanese. Two pages were not blogs at all: one was a site map and another carried banner ads alone. During the period of study, 13 of these blogs became inactive. Some were abandoned by their authors and others were unavailable at their published location. Over time, four of these blogs resurfaced at different locations. These unpredictable behaviors illustrate that even the popular blogs are not immune to incidents of unexpected change.

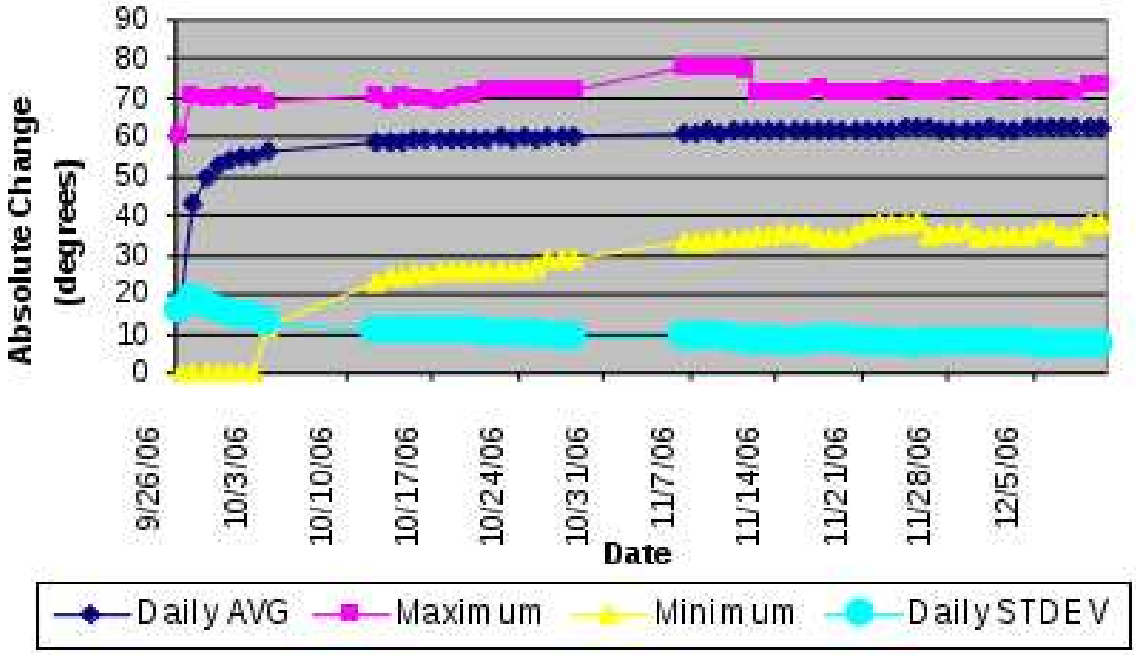

Fig. 1. Absolute Change in Blogs.

pletely dissimilar page. 0 degrees indicates a page with no detected change. The first algorithm compares the base-line sample (the sample cached on September 25, 2006) to each subsequent sample. This measures the evolution of a blog's absolute change to examine the long-term behavior of the blog. The second algorithm compares each daily sample to the prior day's sample, measuring the day-to-day relative changes to examine the blog's short-term behavior. These analyses complement each other, by revealing that blogs follow characteristic patterns of change, both from long-term and short-term perspectives. Figure 1 shows that blogs follow an asymptotic behavior. As time progresses, the absolute change approaches a "stable state" of about 62 degrees of change (absolute change $\mu = 62.37, \sigma = 6.36$). Interestingly, this value is not the full 90 degrees. I attribute this ceiling to the fact that blogs organize their content with fixed templates. While the blog entries change frequently, the templates are

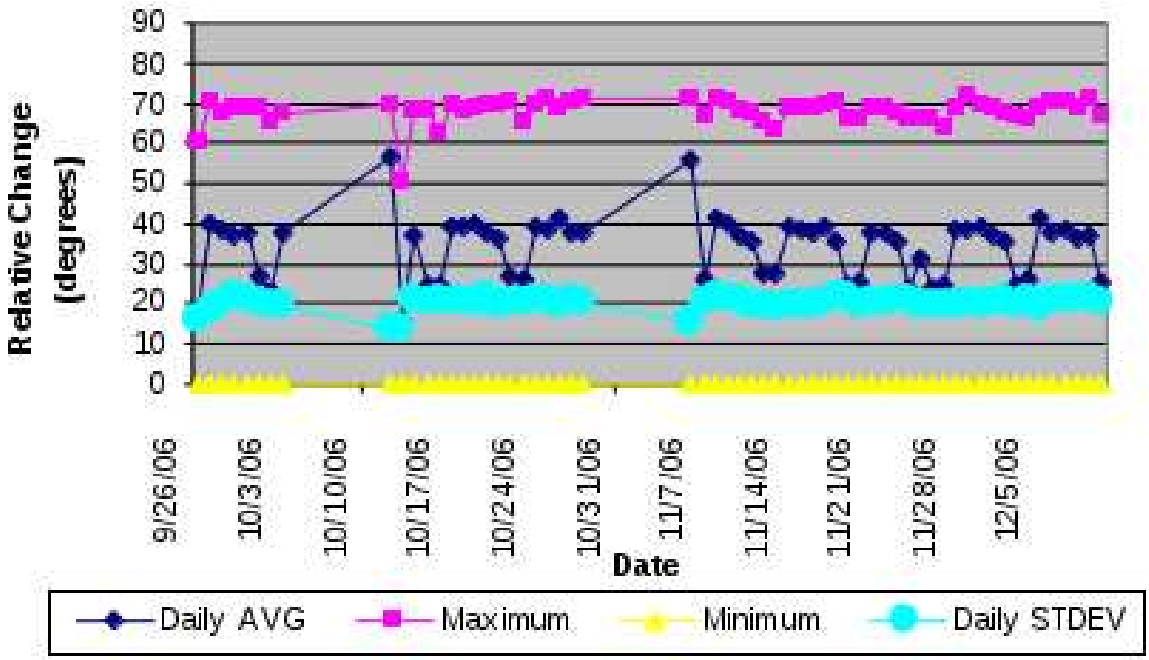

Fig. 2. Relative Change in Blogs.

relatively static. An analysis of the individual pages confirms that the templates vary slightly over time. Changes to the templates are manifested as small oscillations in the blogs "stable state". The "hump" on the maximum absolute change in Figure 1 provides evidence to support this hypothesis. This hump was caused by a blog, www.thoughtmechanics.com, adding a column on November 6 to its template. When the column was removed five days later, the absolute change returned to a stable state. Overall about 25 degrees of the page weight can be attributed to the template ($\mu = 24.92, \sigma = 6.34$) regardless of other factors. Figure 2 reveals that blogs follow a weekly activity cycle. Two spikes are artifacts caused by hardware difficulties with the server responsible for caching, the relative change clearly displays a recurring pattern where activity dips on the weekends. Figure 3 highlights this pattern by presenting the significant difference ($p < 0.01$) between changes effected on weekdays

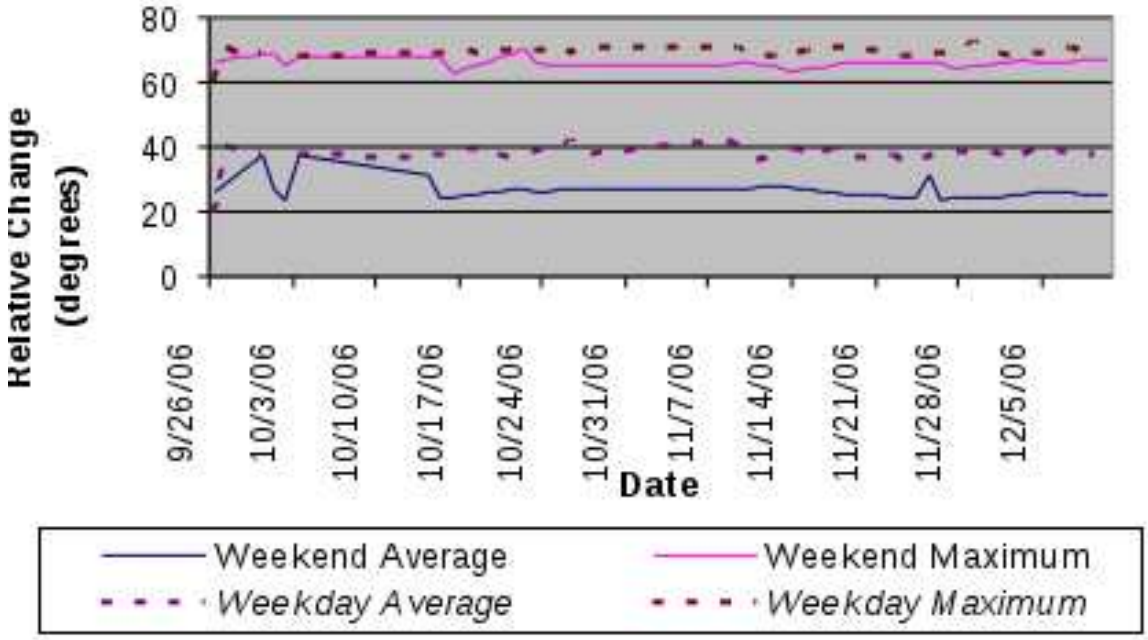

Fig. 3. Change by Day of Week.

($\mu = 38.13, \sigma = 1.57$) and those on weekends and holidays ($\mu = 26.05, \sigma = 2.06$). This study shows that one class of highly dynamic documents, blogs, follows characteristic patterns of change. Moreover, the pattern of change in blogs reflects the differences of human activity cycles in weekdays and weekends. This facilitates the modeling of temporal expectations for web page changes, identifying expected and unexpected changes and, possibly, predicting these changes.

B. Personal Distributed Collections Among Social News Users

As the successor to the PathManager [5], DCM was motivated not only by observations that the fluidity of web pages leads to collections becoming stale and requiring revisions and updates [6], but also by observations that the web as a whole was changing and that assumptions made by PathManager may no longer be valid [7]. Unlike

PathManager, which was focused on maintaining a path in Walden's Paths, DCM is more general and supports other forms of web-based collections, such as bookmark lists and web resource guides.

While a path was a well-defined system artifact produced by Walden's Paths, a general web-based collection, as DCM envisions, is a poorly-defined social artifact. This ambiguity in what a web-based collection could be necessitates an inquiry into what the collections that people are creating are like. Additionally, informal discussions with colleagues raised the question that people may not be creating collections of web pages and are instead relying on recollections and search to re-find previously found web pages. This raised the question of "What value was a system to manage collections of web pages if no one was creating them?". At the time PathManager was created, options, such as social news sites, like reddit and digg, or social bookmarking like del.icio.us, didn't exist. Now that there are options other than plain websites, bookmark files, or recollection, are the user issues that PathManager originally attempted to resolve still relevant? Or, do the social aspects even matter? Lastly, are these collections purely private or do they play a social role?

Beyond these broad motivational questions there also were technical questions that needed to be addressed. Subscription technologies, like RSS, are often seen as a solution to a user staying updated on sites they are interested in, but do they actually improve the problem of staying up-to-date? Does the content-only model of RSS ignore important aspects of a page such as presentation, or interaction? And, are subscription based collections any easier to maintain?

To understand these questions, I conducted an online survey of potential users. From their responses I will show that people do create collections of web pages, that they use a variety of technologies, including RSS, and that the existing tools are inadequate. I will also show that the collections being created, even without

social technologies, often serve a social purpose. Finally, I will show that users are primarily concerned about textual content in their collections, and, despite its focus on textual content, the lack of intelligence in subscription aggregators makes users of subscription-based technologies more likely to be lost in a sea of information.

1. Methodology

For my survey, I used a web-based survey system. I arranged my questions into five sections. First I asked demographic information. The second section focused on personal web-based collections. Here, questions were asked about who used their collections, the tools they used, and how important their collections were to them. The third section delved deeper into the management of collections, asking questions about the kinds of sites in their collections, the kind of changes they care about and their experiences in maintaining these collections. Fourth, I switched specifically to subscription technologies and their likes and dislikes regarding them. The fifth section asks users to identify features they'd like to see in DCM and how likely they were to use a system like DCM for maintaining their collections.

In order to promote the survey I solicited participants through mailing lists and social networks. In particular, I advertised on my lab's mailing list, a departmental list for graduate students and on three social networks – Twitter, Facebook, and reddit. The survey was conducted over a two week period in December 2009.

a. Demographics

I received 125 responses for the survey. 41.6% of the respondents were undergraduate students, 28% were graduate students, while the remaining 30.4% were not students. Ages of respondents ranged from 18 to 52 with the average age of respondents being 25.27. 80 users came from a computing and information sciences background. 12 from

a science background, 10 from a liberal arts and social science background, 8 from engineering, and 1 from education. respondents came from a wide range of localities. North America comprised the majority with with 75 respondents. Additionally, I had 19 Europeans, 6 from Australia and New Zealand, 6 Asians, 2 Middle Easterners and 2 South Americans respond.

2. Results

As discussed previously, I asked questions in roughly four areas: collection usage, management techniques, subscription technologies, and desired features. Statistical analyses were performed using R and gretl. All probabilities, unless otherwise noted, were results of n-way analysis of variance using a linear model with factor interaction taken in to account.

a. Collection Usage

Several questions asked by my survey focused on the usage of collections of web pages. The first question was if they had collections of web-pages. 45.6% of respondents reporting having a collection of web sites. However, an additional 15.2% indicated later in the survey that they did maintain a collection when I asked about more specific kinds of collections, which implies a true total rate of collection creation at 60.8%. Of those who have collections, only 4.5% reported that they never revisit their collections, while 80.3% revisit their collections daily.

The next question was if collections were private or if they were shared. Only 22.81% of the respondents indicated that someone other than themselves used their collections. 53.85% of respondents who shared their collections of web sites did so with family. These respondents created collections that tended to change more often than collections created by people not sharing with their family members ($p = 0.05$).

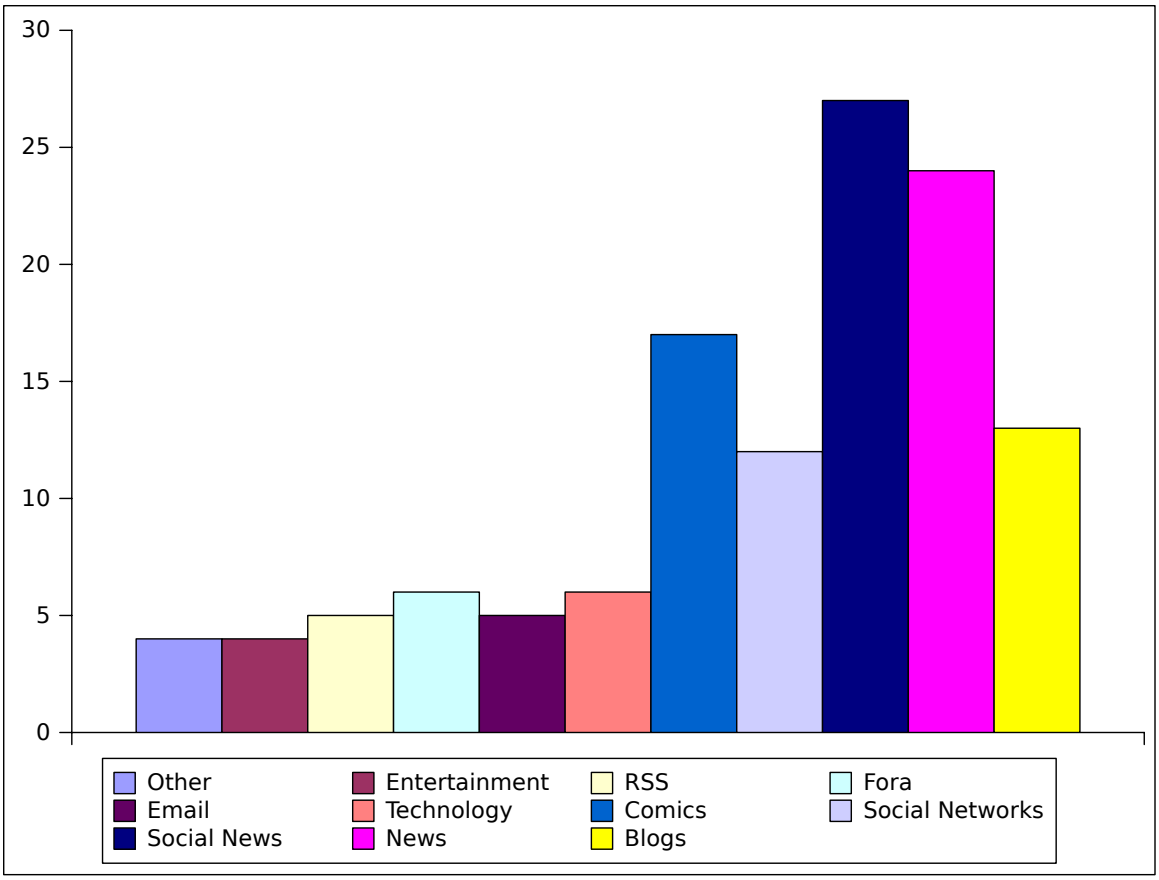

Fig. 4. Types of Sites in User Collections.

23.08% of those who shared, were sharing their collections with friends. They tended to lose track of their collections more often than respondents who weren't sharing with friends ($p = 0.08$). 69.23% indicated small groups of people either in organizations, a work environment, or in a academic project group. There respondents created more frequently changing collections than people whose collections were not being used by a group ($p = 0.07$). Likewise people who created both collections that were used by their family and in a professional/academic setting tended to have the most frequently changing collections ($p = 0.04$).

As Figure 4 shows, when I asked what type of sites people were interested in for their collections, social news sites and traditional news sites dominate the kind of

sites that respondents keep in their collection of web sites. Comics come in at third while blogs and social networks were cited the fourth and fifth most frequently.

b. Collection Management Techniques

Another area of interest was what tools people were using to maintain their collection. Every respondent except for one reported using some sort of tool for maintaining their personal collections. Traditional bookmark usage was common, with 85.45% of respondents using them. However, despite the fact that the majority of respondents (57.14%) were consumers of social news and bookmarking sites, only 23.64% of respondents were actually using social news or social bookmarking sites to maintain their collections. 12.73% of respondents were using a subscription technology like RSS and 11% were using other kinds of web pages (like wikis or hand-written HTML) to maintain their collection. I found that for certain factors, the kind of tool was a statistically significant detriment to the respondent using the tool. Respondents using bookmarks found it more difficult to maintain their collections than respondents who didn't ($p = 0.02$). Respondents using no tools ($p = 0.03$), their history mechanism ($p = 0.08$), or their email ($p = 0.08$) to maintain their collections perceived them changing more dramatically than others.

Of the users who used a subscription-based technology, all of them also used bookmarks, and 14.29% of them also used some sort of web site. 52.17% of bookmark users used another technology.

For types of change my results appear contradictory to speculations made by others in the literature. Content changes made up the vast majority of changes people were interested in. 89.5% of respondents indicated "content" as an aspect of change they were interested in. The second-highest aspect was "visual" with only 5.08% interested. However, I suspect that some of "content" as defined by the respondents

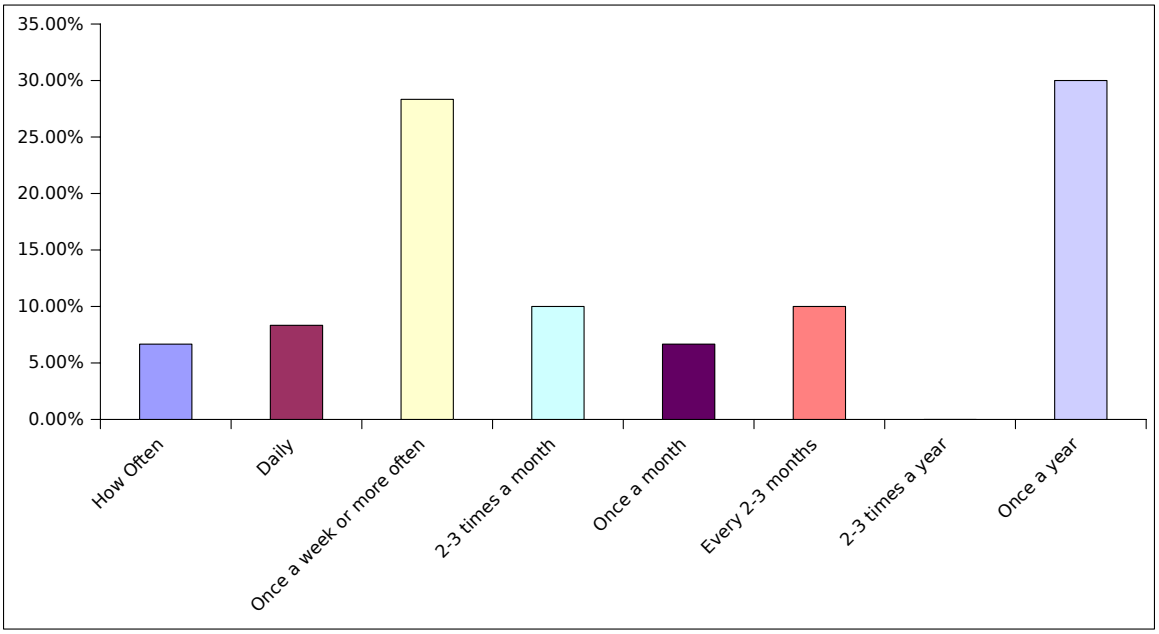

Fig. 5. Relative Frequency in Which Users Lose Track of Their Collections.

still included imagery, particularly since comic sites showed such a frequent occurrence in respondent collections.

Respondents were asked "How often would you say that you lose track of sites in your collection?" They were given the option of daily, once a week or more often, "2-3 times a month", "once a month", "every 2-3 months", "2-3 times a year", "once a year", "rarely" or "never". As Figure 5 shows, I found a bimodal distribution with means at "2-3 times a month" and "never". However, I was not able to correlate the bimodality of my results to any other factor.

I performed a Pearson's coefficient calculation between each pair of questions. From these coefficients I was able to find a number of correlations between factors dealing with collections. People who create work collections were found to have less dramatic changes than other kinds of collections ($p = 0.06$). The more important a collection was to a respondent, the more time they spent maintaining it ($p = 0.09$)

and the more difficulty they had in keeping track of it ($p = 0.11$). Collections that were revisited more often were also more difficult to maintain ($p = 0.10$). Difficult to maintain collections took more time to maintain ($p < 0.01$). Subscription-based collections took more time to maintain than non-subscription technologies ($p = 0.02$).

c. Subscription Technologies

When respondents were asked what they liked and disliked about subscription technologies, 86.2% of respondents had the same like – consolidation of several sites content into one easy, quick place to read everything. However, four major kinds of dislikes were found. 37.5% of them said that the pace of updates caused information overload and that they need some kind of filtering method. 33.3% complained that the subscription feeds were often only a subset of the content of the site. Some feeds would miss items, some wouldn't have consistent metadata, others wouldn't have the entire article text, and some wouldn't provide locations of relevant images. 12.5% found the selection of sites to be publishing feeds to be sub-par or limited and finally, 8.3% found the interfaces of the readers themselves to be inadequate.

d. Desired Features

Finally, I asked users what features they were interested in for a system for managing their collections of web pages. 36 users provided substantive answers. Of those 36, 14 indicated various social web features like sharing, voting, tagging, and recommendation. 12 indicated that they wanted a system that was easy and simple. 7 users wanted to be automatically informed of updates, 6 wanted categorization, 5 wanted to be able to define filters or priorities to limit information from sources they were less interested in, 4 wanted to be able to easily view collection members from inside the system.

C. Characteristics of Bookmark Usage Among Graduate Students

In order to dig deeper into my results from my broad survey of potential users of a collection management system, I conducted interviews with twelve students enrolled in a graduate course on digital libraries. Half of the students were male, the other half were female. All of the students were between 22 and 28 years old. The students were asked to bring in a collection of URLs such as, bookmarks, a resource list or a set of RSS feeds, to look over and answer questions about during the interview.

While one user did bring in a set of RSS feeds, the remainder of the users brought bookmark files. No interviewee created a hierarchy or categorization scheme more than two levels deep. Regardless of whether the users used one or two levels of hierarchy, one of the levels was genre-based. The second level had a variety of uses from time-based grouping, to (sub/super)-genre classifications, to grouping sites. This genre classification has a greater importance as users specified that they treat sites differently based on genres. Additionally, two users indicated that maintaining hierarchies proved to be too time consuming and that they eventually abandoned all categorization.

While one user did use Google search as their only means of revisitations, most users did use their bookmarks. However, they did not always recognize that they did. In fact the most common way to re-access their bookmarks was using the URL bar to search their own bookmarks and browser history. Additionally, two users relied on the recently visited feature in Google Chrome to revisit sites that they had bookmarked. Several people used the bookmarks bar in the browser as a quick means to revisit frequently visited sites. Many people, particularly the bookmark bar users, indicated that frequently visited sites were not categorized.

Sites of interest reflected the previous study with web comics, entertainment

sites, news, mail, social sites, and blogs being the commonly cited kinds of pages. Additionally, due to the fact that all interviewees were students, all of them stated that not only were course-related websites important to them, but that they visited them at least every other day. They also indicated that course sites required frequent attention. Due to the dominance of computer science backgrounds, most users also had technical documentation in their collections. These sites were visited more infrequently than other kinds of sites, expected to be static and changes were viewed as very problematic.

In general, revisitation of sites that were bookmarked either fell into frequent revisitiation (from every other day to many times a day) or rare to never. The sites that were rarely to never revisited tended to have been bookmarked for a particular past task or as a reminder to read later. Despite these good intentions, many users admitted that they had forgot those links were there and that they need to "clean up" their bookmarks. One factor that many users indicated that informed their use of bookmarks were the tasks they were engaged in. A second factor was dynamicism. The more a site was expected to change, the more the interviewees cared about how the page changed.

The kinds of changes they cared about were primarily textual. The notable exceptions were images that contained information that was essential for the meaning of the site. Examples of this were given by two users in particular. One of them was interested in role-playing games that centered around collectible cards. For these games there exists sites discussing strategies and the various available sets of cards. When an image changed on these sites it could have broad implications to game mechanics and the usage of particular cards. The second user was very involved in certain fan-fiction communities online. To him, an image change in a story of fan-fiction could indicate a refinement of the author's vision for a piece of fan-fiction

which could alter his understanding of the story he had read previously.

In summary, I confirmed that my reading of the broad survey results corresponded to what people were thinking and helped to explain the vague, "content only" answer that survey participants indicated can be explained to mean that imagery counts as content when the image is integral in determining the meaning of the entire page.

CHAPTER VI

THE DISTRIBUTED COLLECTION MANAGER

To address the problems of collection management, I created the Distributed Collection Manager (DCM). Currently, DCM provides a server side repository of caches of pages that is updated every four hours. Additionally, DCM contains a framework to build client-side applications that leverage this repository. The goal of this work is to provide an environment to analyze and experiment with maintaining collections. This environment consists of an integrated client application, metrics of change, and visualizations of those metrics.

A. General Architecture

DCM is envisioned as an ecosystem of tools tied by a common library, repository and database. The library leverages the Apache Cayenne ORM to provide a common set of classes and data structures backed by a MySQL database. The database contains all of the metadata, task records and system outputs with the exception of actual caches of web pages and images. Each cache and image is stored instead in a subversion repository. Each HTML page or image's URL is hashed to the hexadecimal equivalent of the URL and a file is created in the subversion repository with the hash as its name. New versions of the same page or image are stored as new revisions to the files in the repository. This allows storage of only the differences between files and quick and logical retrieval of revisions.

B. Nona

The ancient Romans revered a set of three deities that they called the Parcae, or the fates. Each of these goddesses controlled one aspect of life. Nona, was at first a goddess of pregnancy and the manifestation of the ninth month of pregnancy. She, in the cloth-making analogy of the fates, was the spinner of the thread of life. Likewise, in DCM, Nona is the system that draws pages from the web and places them into the repository.

Nona is a server process that four times a day queries the database for all pages to be cached and schedules caches to be taken with the requested periodicity using cron-like syntax. Each scheduling thread waits until its time arrives, whereupon Nona retrieves the current version of the web resource. The source is then parsed for images which are also retrieved. Finally, Nona stores the page and images as new revisions in the repository.

C. Decima

The second of the Parcae is Decima. While Nona was a goddess of pregnancy, Decima is a goddess of childbirth. Decima's job was to usher the child into the world and determine its lifespan. Using the cloth-making analogy, Decima is the one who measures the thread of life. For DCM, Decima is a modular feature extraction system that is used to measure important metadata from a cache for use in analysis.

Like Nona, Decima is a server process, unlike Nona, Decima runs once a day. Daily, Decima wakes from sleep and retrieves the list of all caches missing results for scheduled features. For each pair of cache and unextracted feature, Decima adds an extraction task to a thread pool. The features are specified by a custom format, a feature archive (FAR). FARs can be a directory or an archive, using Zip compression. A

FAR contains three parts, first a custom XML property file, called a feature description language file (fdl), that encodes a chain of filters for a Mallet feature extractor [61]. Any custom processes to be executed are contained in a classes directory in the FAR file and any third-party libraries needed for the FAR to run are included in a lib directory. Using the 3rd-party JCL reflection tools, I automatically add needed classes and library dependencies for each step in the fdl file. This constructs a Mallet processing chain that is then run on the cache for the task. The result is then stored in to the database for use by the analysis system.

Currently Decima has six features implemented – a term vector generator with stopword removal and stemming [1], a Flesch-Kincaid readability index [63], a structural count based on Johnson's algorithm used in PathMananger [6], punctuation counts, a count of unique outgoing links, and a n-gram (for n of 2,3,4,5) generator [64].

D. Phaeton

As a bit of a linguistic quirk, Decima is also a Latinized spelling of Dejima, a man-made island in Nagasaki bay that was home to first Portuguese and later Dutch trading contingents during the isolationist period of the Japanese Shogunate. In 1808, a British warship named the H.M.S. Phaeton entered Nagasaki bay and held the city, and particularly the Dutch trading contingent on Dejima at gunpoint to extort goods and damage the trade routes of the now Napoleon-controlled Dutch. After the discovery of the importance of page images from the study on social news users, I set out to supplement Decima with a system specifically for extracting features from images. Much like how the H.M.S. Phaeton harassed Dejima island by unexpectedly arriving

[1] Stemming is accomplished using the Snowball English Stemmer[62]

in Nagasaki bay, Phaeton is late additional to DCM that was initially unexpected.

Phaeton is now the third server-side process and works much in the same way as Decima. The major difference is the Phaeton Image Feature (PIF) files, are not controlled by a property file, but instead contain a folder containing classes that implement the PhaetonFeature interface. A PhaetonFeature currently assumes a default constructor and a single public function that accepts a BufferedImage as a parameter and returns a Mallet FeatureVector. Due to the rough current nature of Phaeton, it was not used for analysis in the following chapters.

E. Morta

The final of the Parcae is Morta, the goddess of pain and death. In the cloth-making analogy, Morta is the one who cuts the thread of life when the end is reached. For DCM, Morta is the analysis system. Morta is the system responsible for taking the features extracted by Decima and generating a set of error measurements relating to the Kalman filter predictions.

Morta was originally created as a single system but as work progressed I split Morta into three parts: Premorta that prepares tasks, OctMorta that runs the tasks, and Postmorta that stores the results back into the database.

1. Premorta

Premorta is a server-side application that once a day builds several text files that OctMorta needs to perform its analysis. The primary file is an Octave-format sparse matrix file containing the full feature vectors for each cache in the analysis period [65]. The second is a file that contains the mapping from features to rows. The third is a file that contains the mappings from dates to columns. These files are

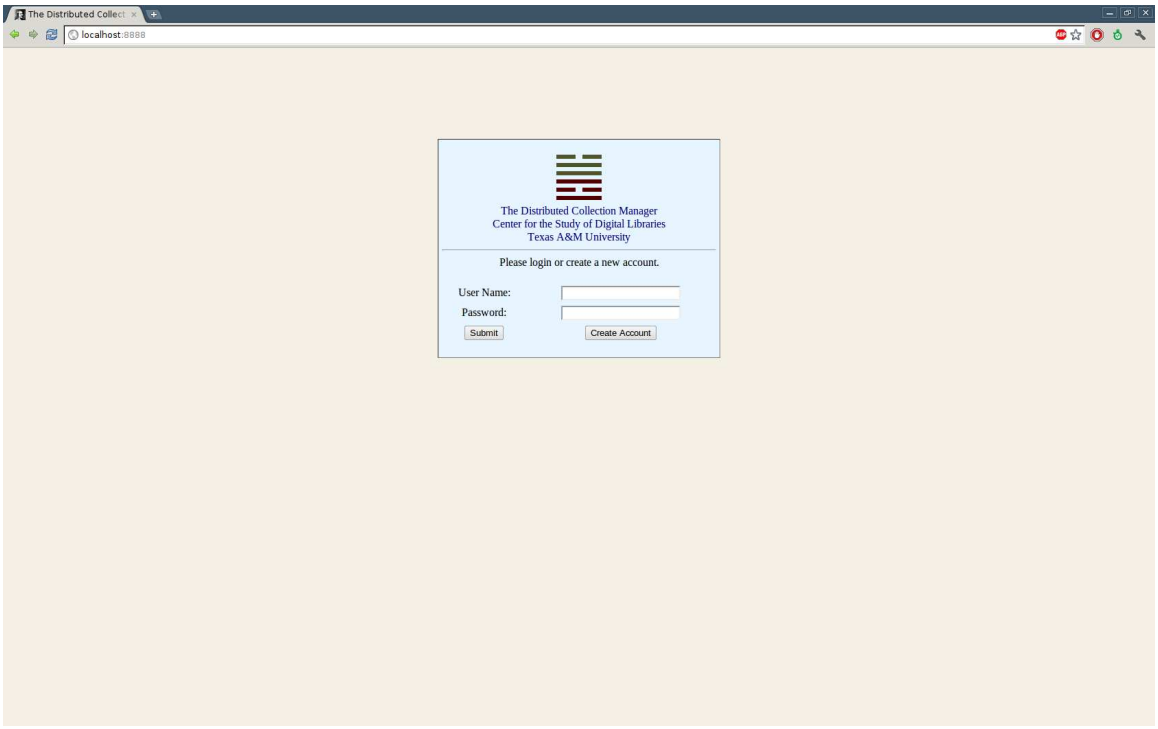

Fig. 6. Hannah 2: Login Screen.

placed in a directory on DCM's server to await transfer by OctMorta to the Brazos Supercomputer [2] at Texas A&M University.

2. OctMorta

OctMorta consists of three parts. The first part is a bash script that synchronizes the input files on the main DCM server with the versions on the Brazos Supercomputer at Texas A&M. Then another bash script creates a list of analysis tasks without up-to-date results and submits them in groups of eight [3] to the scheduling system on Brazos. The second portion is a batch script executed when a scheduled task is run that

[2] I acknowledge the Texas A&M University Brazos HPC cluster that contributed to the research reported here. (http://brazos.tamu.edu)

[3] Tasks are grouped into sets of eight to take advantage of all eight cores on Brazos compute nodes.

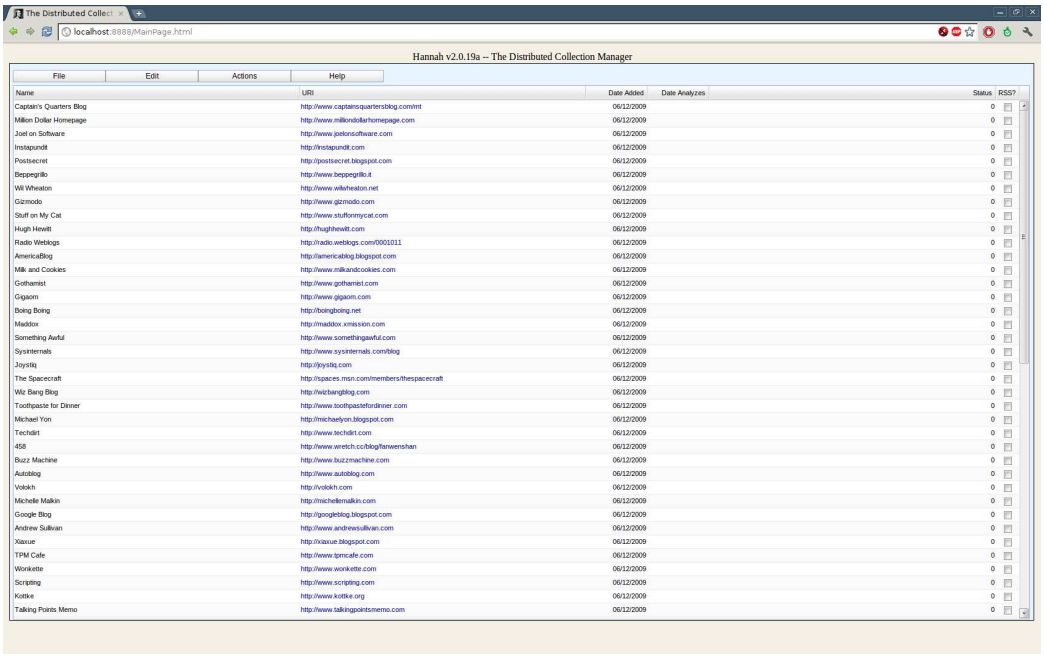

Fig. 7. Hannah 2: Main Screen.

configures the Octave bash and then launches a program in Octave. The program first loads the matrix, then determines a covariance matrix in order to perform principle component analysis, and then finally steps through the dimensionally reduced data predicting the current state and then adding the observed results for the current state as a new observation to the Kalman filter. The error between the prediction and the next observation are calculated and stored into an output file. After the second part completes, the third part, a final bash script, then synchronizes the error lists back to the DCM server for re-incorporation into the DCM database by Postmorta.

3. Postmorta

Finally, once a day, Postmorta checks for new results on the DCM server from Oct-Morta and then enters into the database each cache's result along with a running mean and standard deviation of the error between observed and predicted values.

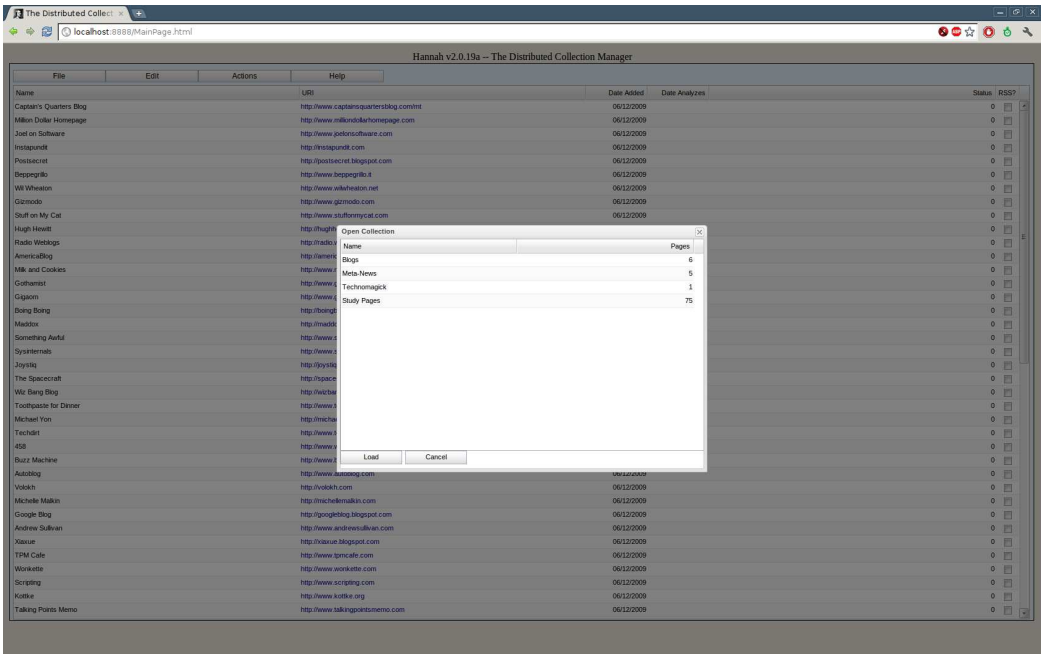

Fig. 8. Hannah 2: Open Dialog.

F. Hannah 2

As an initial interface into the DCM, I built Hannah 2, named after the pioneer of primary school libraries, Hannah Logasa. Hannah 2 is a web-application using the Google Web Toolkit for building collections and scheduling caching, feature-extraction, and analysis. DCM is an expert tool designed for a power-user wishing to have total control of how DCM handles their collection.

Hannah 2 allows users to create multiple collections each with their own features to be analyzed and their own caching periodicity. Users can explore each cache in the repository and every extracted feature. Finally, they can view the results from Morta and provide feedback to help tune the sensitivity factor for determining if a cache represents an expected or unexpected change.

As Figure 6 shows, a curator using Hannah 2 must first login to an individual account or create a new account. The curator is then presented with a main screen as

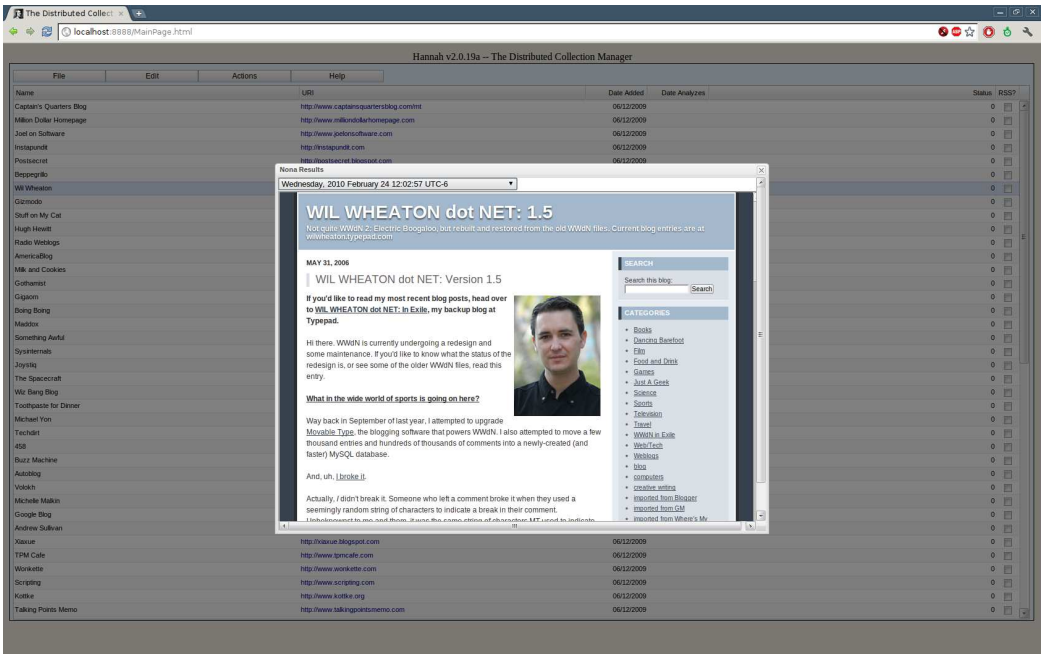

Fig. 9. Hannah 2: Cache View Dialog.

show in Figure 7. This screen allows her to create, edit, load, and import collections as well as provides a place to schedule tasks for DCM back-end systems as well as view results from those systems. When a curator selects open in the menu they are presented with a listing of collections they own along with collection-level metadata to help pick one to load. This can be seen in Figure 8. Caches collected by Nona can be browsed in Hannah 2 as seen in Figure 9. For feature extractors, as shown in Figure 10, Hannah 2 provides a graph visualization indicating the progress over the collection. Morta scheduling and results are handled in a seperate screen as seen in Figure 11 this allows multiple different analysis periods with different steps to be scheduled for each collection. The results produced by Morta, like Decima' results, are also viewed using a graph visualization. An example of this can be seen in Figure 12.

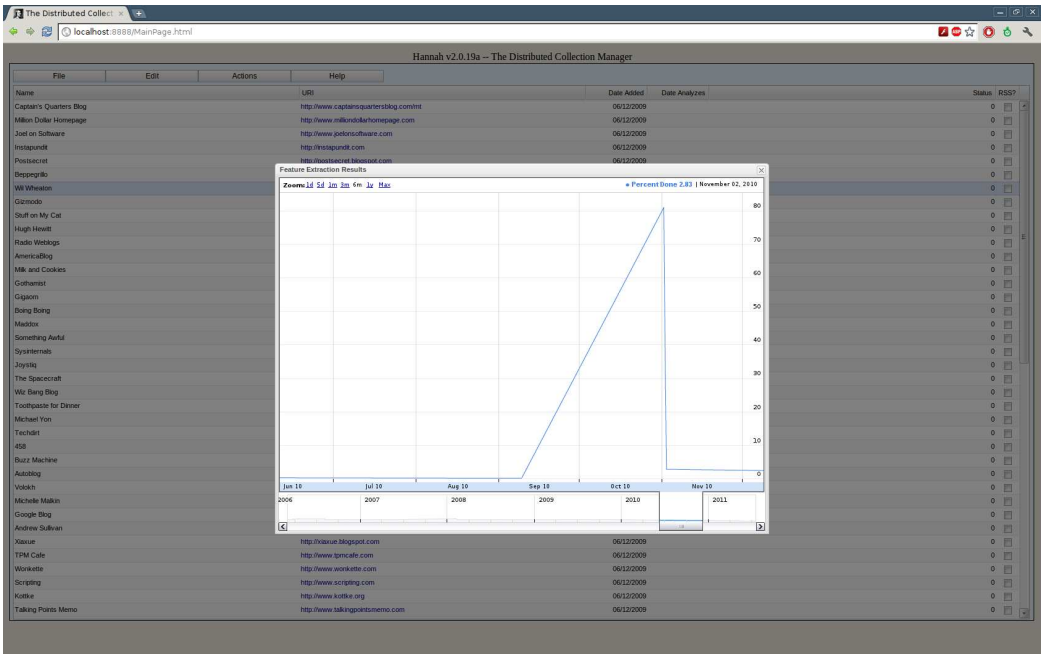

Fig. 10. Hannah 2: Feature Extraction Status Dialog.

G. Ananake

The next portion of my system, Ananake, is a specialized web-crawler that searches a single domain for outgoing links. These links are used to build a collection for DCM. The envisioned use case is to help site administrators ensure that the links they point to are still the same as what the author of the page containing the link intended. For Ananake, a domain consists of all sites reachable from a predefined starting point such that all of the pages are in the same top-level domain without traveling a link that points to another top-level domain. The main limitation of this approach is the crawler can only find a single semi-connected component rooted at the starting point. If there are other components that are not reachable by internal links from the starting point, they will not be crawled. For smaller resulting collections, Hannah 2 can be used to perform maintenance tasks. However, larger collections may not be easily presented in Hannah 2 and require another tool, Ianus, instead.

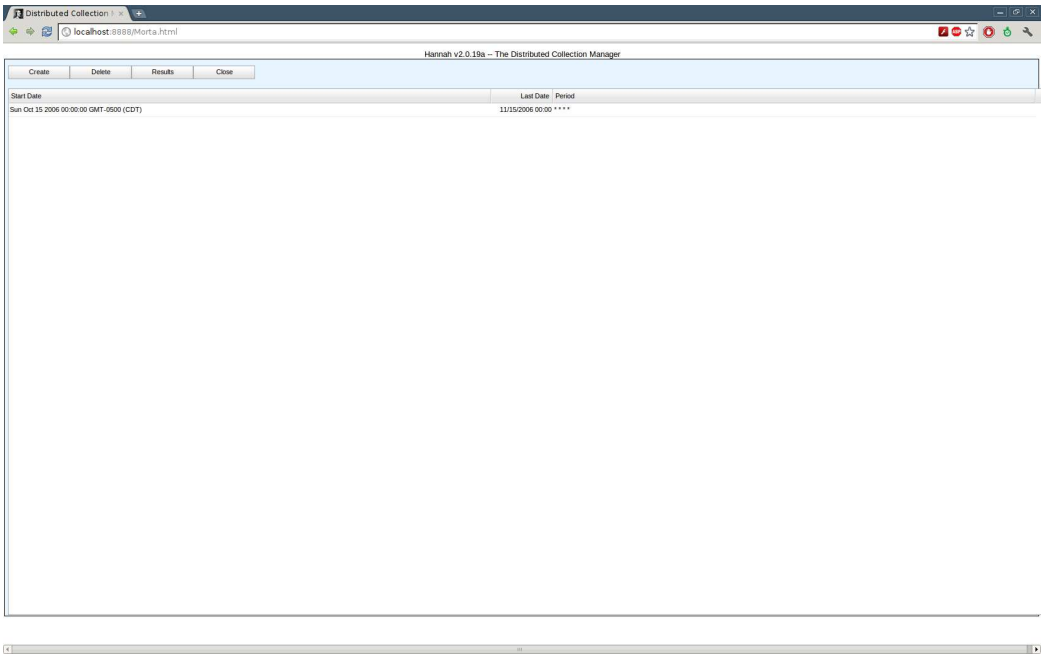

Fig. 11. Hannah 2: Analysis Scheduling Screen.

H. Ianus

Ianus is a prototype problem notification system design for large collections. Ianus is meant to easily allow maintainers to be notified when an unexpected change occurs in their system. Ianus acts by checking the results of scheduled analysis tasks and emailing the collection owner a summary of the collection's status including potential problem pages.

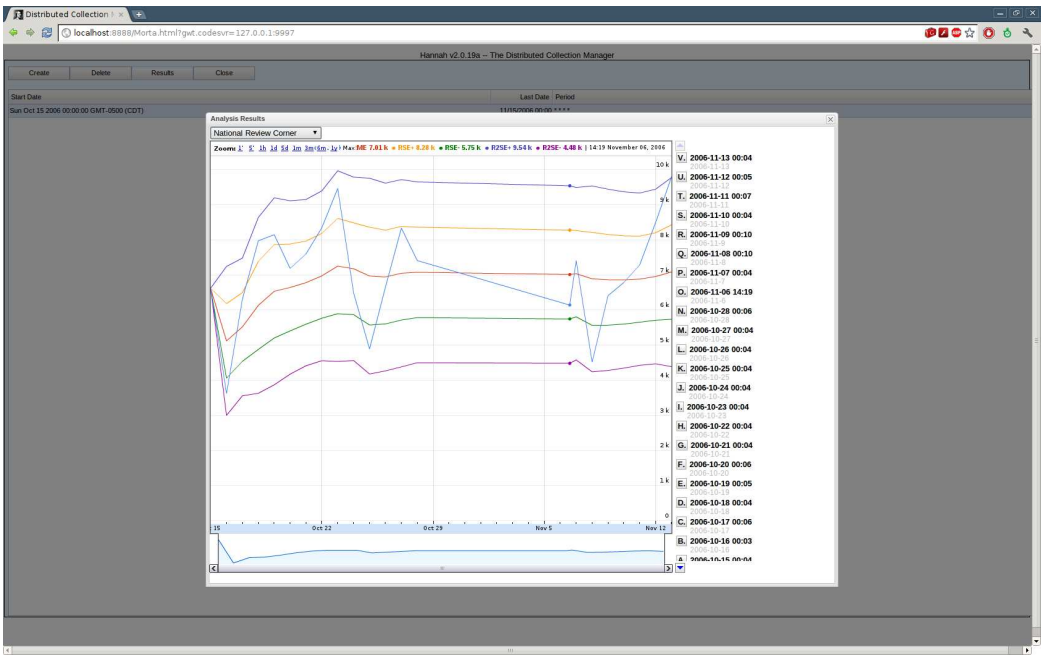

Fig. 12. Hannah 2: Analysis Results Dialog.

CHAPTER VII

KALMAN FILTER PAGE CHANGE ANALYSIS

> To succeed, planning alone is insufficient. One must improvise as well.
>
> Isaac Asimov, *Foundation*

In the case of modeling change in web pages, I am formulating the problem as an analogy to a point moving in a high dimensional space. As mentioned previously, while it is not clear that content changes occur in a linear fashion, simple linear models often perform well and provide a good staring point. Therefore, I selected Kalman filters to model how a blog changes with time. In order to find unexpected change, I train a Kalman filter with the current history of a web resource and use the filter to predict the next expected version. I then use a measure of distance from this expected version to the observed, true next version to calculate an error. This error is then converted in to an instantaneous normalized score compared to the running mean and standard deviation of error. These normalized scores are then used to determine if a change was expected or not. A discussion of various techniques that I tested to make that determination is given in the next chapter.

The basic underlying assumption is that when a collection maintainer revisits a web resource in their collection, they have an immediate visceral response that draws upon their experience with the web resource in the past. This response informs the collection maintainer whether a deeper evaluation is needed or if the page is still what they expect it to be. From my studies with users of social news, I had concluded that textual and imagery changes are of primary importance to users in making these visceral decisions. Therefore by focusing on features that highlight these kinds of changes, I can emulate part of the process the collection maintainer undergoes. Since this response is tied to the past, any method modelling expectation needs to take into

account the history the user has had with the page. In the past a similar problem was found with radar systems. When a radar operator was tracking an object across the screen, the discrete steps from each sweep led to ambiguous situations where it became questionable if the artifact they saw in the new step was the same object as the one they saw before. The solution to this problem was to use the history of the object's motion to try and predict were the object should be with a reasonable amount of error. This generic description of the radar tracking problem resembles the problem of tracking web-pages. While none of the techniques I used are necessarily tied to features of blogs, other than the assumption of change being expected, the results as follows are on the collection of blogs mentioned in prior sections.

A. Feature Selection

When revisiting a site, our surveyed social media users and interviewed graduate students indicated that their primary concerns when re-evaluating resources were in the areas of content, topic, target audience, and authorship. Because of these concerns I sought features that attempt to address these issues. While my visual similarity features are currently constrained to those focusing on color schemes, I have more directly addressed the other stated areas. For topic I included all 2-,3- grams in a document and the term vectors that have had stopwords removed and have been stemmed using Porter's Snowball stemmer [62]. For the target audience, I perform a Flesch-Kincaid readability index [63]. For authorship I perform a punctuation analysis, and a structural analysis. Each of these features is used to create an ensemble feature vector that is then reduced using principal component analysis before being processed by my Kalman Filter.

B. Change in Blogs Ground Truth Corpus

In order to evaluate my methods, I created a corpus consisting of daily caches during the two month period surrounding the 2006 United States Congressional mid-term election. This period was picked both for its completeness, as there were no system or network outages, and since it was a political charged time, it is expected to produce more interesting results in many of the blogs in the collection. A total of 62 days of 66 blogs were first checked for static instances – days where no addition occurred – this reduced the blogs to 22 interesting days. These caches were then processed with an HTML differencing engine to produce an easier to scan display of what had changed. Each of these caches was shown to three evaluators who selected from a variety of classifications of change to tag the newer version. The options we provided were: no change, change too slow, change too fast, topic shift, abnormal content, and normal change. Due to the size of the collection, evaluators were given as much time as they wanted to spend to tag as many as they could. While differences within a cache were shown consecutively in order of capture date, the order of sets of pages were randomly shuffled. Once a cache had been evaluated by three different judges, it was not shown to future judges. This corpus was then used to quantitatively compare different combinations of features, and outlier detection methods. Additionally, I quantitatively compared the results of my Kalman Filter based analysis to variety of threshold based analysis techniques. A discussion of these results is contained in the following chapter.

CHAPTER VIII

ANALYSIS

The analysis system at the heart of DCM contains three important parts which require an independent analysis for each. The first of these parts is the selection of features. Secondly, I compared different techniques for identifying the outliers. Finally, I compare the overall best feature with the best outlier detector to a cosine-similarity based threshold technique that been used by prior change detection systems.

A. Feature Selection

Traditionally, feature selection is a fine-grained process that, due to the inherent computational complexity of calculating permutations, cannot be done exhaustively. Term vectors, in particular, have the potential to reach impractical running times even with small dictionaries. To avoid calaculating every permutation, there are a variety of techniques to trim the decision space and test only options that are likely to provide an improvement. However, since DCM is built on coarse-grained feature extractors that each produce a number of features and because these features are reduced through principal components analysis, we can forgo the fine-grained analysis of each individual feature in preference of combinations of extractors. While better results may be found when selecting features individually, the results are more prone to overfitting [66] and don't help to provide insights into the usefulness of my coarse-grained extractors as a whole.

From these facts, I proceeded with a naive feature selection of all sixty-three combinations of the six features I currently have implemented. When a page is prepared for analysis by Premorta, task files for each combination of features are generated. These files are then sent to the Brazos Supercomputer for analysis.

From the results of all feature combinations I performed an ANOVA on F-Scores and accuracies produced by each feature combination. Accuracy is defined in equation 8.1. An F-Score is the harmonic mean of percision and recall. The calculation for an F-Score is described in equation 8.2. Percision and recall are defined in equations 8.3 and 8.4 respectively.

$$A = \frac{TP + FP}{TP + TN + FP + FN} \tag{8.1}$$

$$F = 2 * \frac{P * R}{P + R} \tag{8.2}$$

$$P = \frac{TP}{TP + FP} \tag{8.3}$$

$$P = \frac{TP}{TP + FN} \tag{8.4}$$

I determined that none of the features alone or any combination thereof provided a statistically significant improvement over any other. This most likely is explained by the application of principal component analysis performed by Morta before the Kalman filter is executed. Principal component analysis, by projecting the data into a lower dimensionality that maximizes the expression of the data's variance, would explain the wide variance I found for feature-wise analysis. This wide variance would result in a poor separation of feature-based classes thus resulting in needing to accept the null hypothesis (that the features are equivalent) in an ANOVA analysis.

B. Outlier Detection

The output of Morta is a list of how far off each prediction made by the Kalman filter was in compared to the actual observed results. These error measurements when standardized and manually inspected seemed to indicate a correlation between the standard score and the likelihood that they were abnormal pages. To quantify this observation I devised a number of candidate methods to pick out the abnormal pages. These methods can be divided into three categories – heuristic methods, classifiers, and statistical outlier detectors. Each of these categories has several alternate methods while sharing a common perspective on how to treat this data. The theory behind most of these techniques is that users can provide feedback to help improve future attempts to determine if an error measurement is normal or abnormal. For the purpose of testing, I am simulating the case where users always provide feedback. This is done by retraining the the techniques by feeding the results of my ground truth study back into it after classification of a revision.

Due to the nature of results for machine learning type tasks, a single simple measure cannot capture the complexity of the data's results. For the remainder of this section we will present both the overall accuracy of the method and an F-Score, a commonly used technique that reports a balanced combination of precision and recall. Some techniques perform very well in terms of F-Score and some in terms in Accuracy. A few techniques perform well in both measures, albeit not as well as the maximal techniques for each measure independently. Many measures perform poorly on both and are clearly inappropriate for the task at hand. In particular, we have found Spectral Regression Discriminant Analysis [67] and nearest neighbor techniques [68] to be the most interesting for deployment.

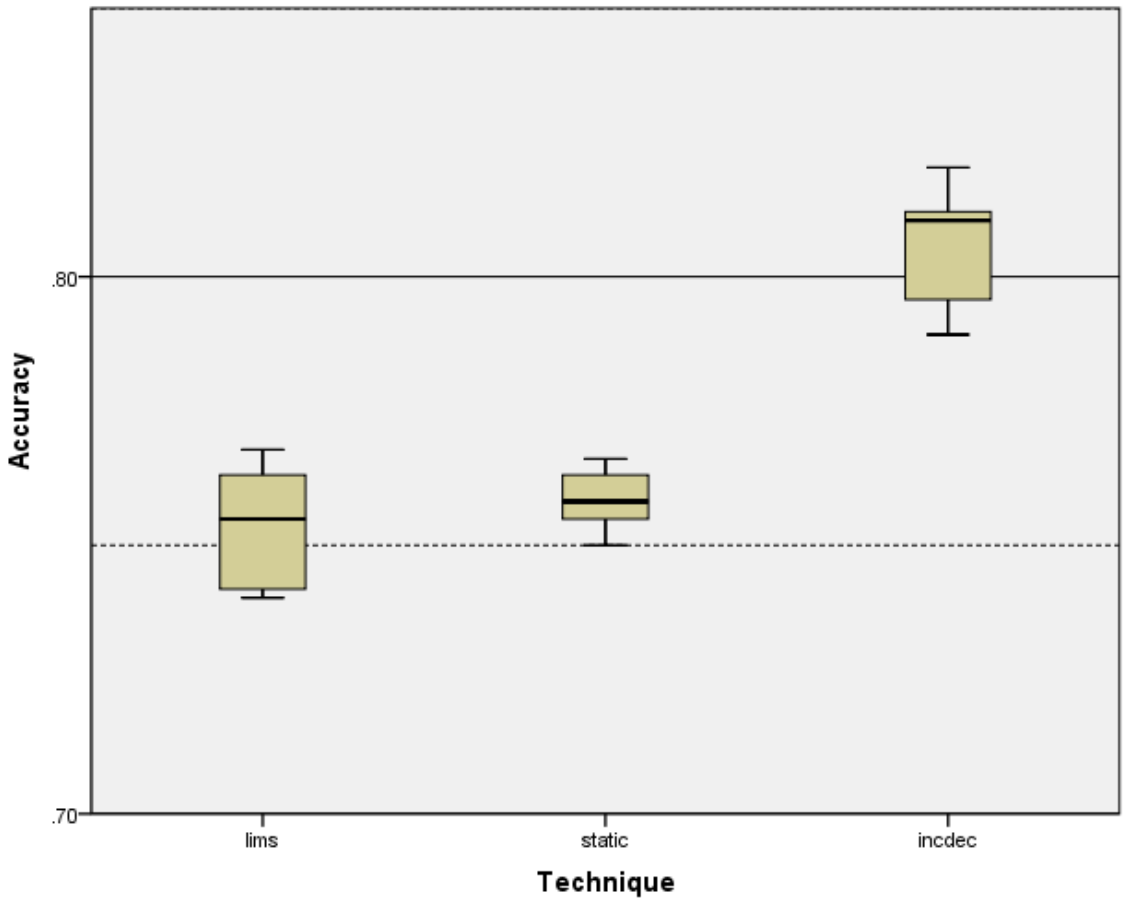

Fig. 13. Boxplot of Accuracies for Heuristic Methods.

1. Heuristic Methods

The heuristic methods are often the simplest and are informed by my observations of the data. Currently, I have tested three different heuristic methods. Results are summarized in Figures 13 and 14 and discussed in the remainder of this subsection.

The first method and by far the simplest is a static error threshold placed at two standard deviations above and below the mean for past errors. If the new revision's prediction error is outside the range between the two thresholds then it is marked as abnormal, if it is between it is normal. To provide an example of how the static

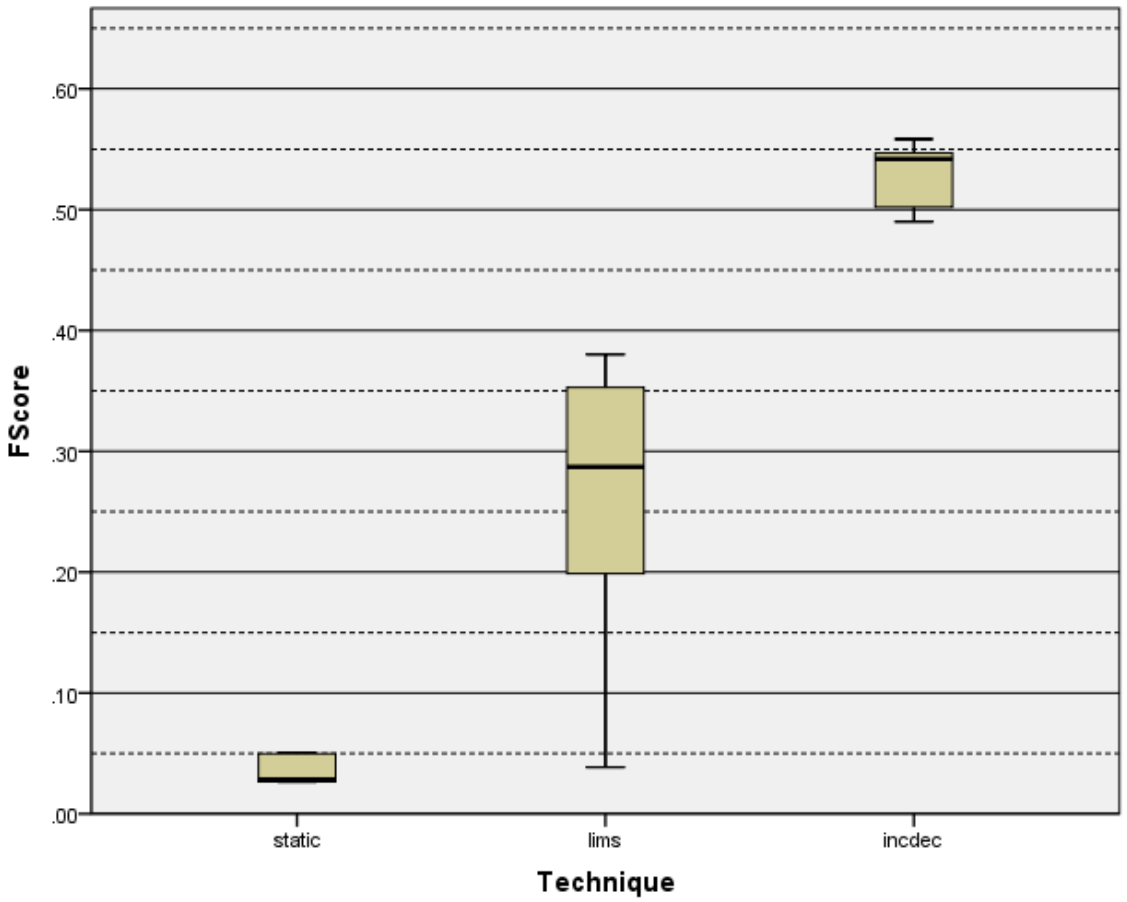

Fig. 14. Boxplot of F-Scores for Heuristic Methods.

technique behaves, please see Figure 15.

The purpose of a static error threshold was more to provide a baseline to compare techniques from the most naive means of analysis – one that does not change in response to feedback. The static technique provided one of the best of overall accuracies (75.86%) at the cost of one of the worst overall F-Scores (.035). This reflects the overall dominance of negative (normal behavior) results in our collection. The static method tended to err on the side of more negative items, thus having a good accuracy (mainly by chance) with such a poor F-Score.

The second method increases the complexity by allowing the bounds to grow

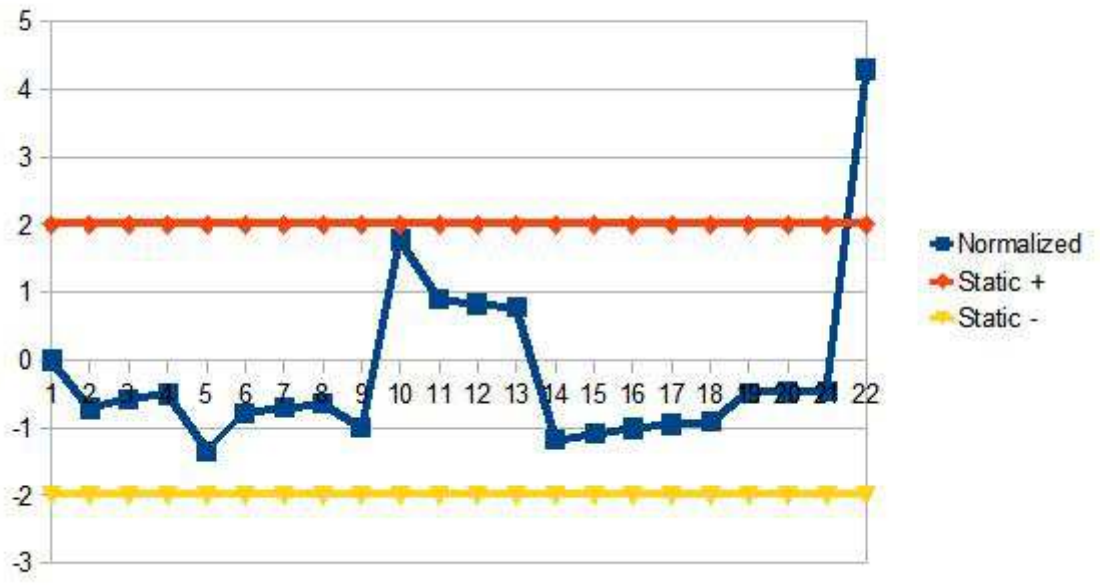

Fig. 15. Static Technique Results on Outer Court Blog.

and shrink using constant steps from a starting point (two standard deviations) in response to false positives and false negatives. In a production system this information would be obtained from user feedback, thus in the case of no feedback it will perform like the previous method, while with perfect feedback it will perform as well as the algorithm can be expected to perform. One thing to note is the threshold above and below the mean change independently of each other. To provide an example of the results from the constant step technique, please see Figure 16.

Allowing constant steps has the effect of dampening the effect of highly abnormal results – a revision with error far outside other revisions, or an abnormal cache near the mean. This can be viewed as a conservative heuristic. Surprisingly, this conservative method obtained the best over all accuracy (80.67%) and a slightly better than expected random classification F-Score (.531). Unfortunately in cases where the misclassified cache is distant from the threshold, repeated similar caches may still be

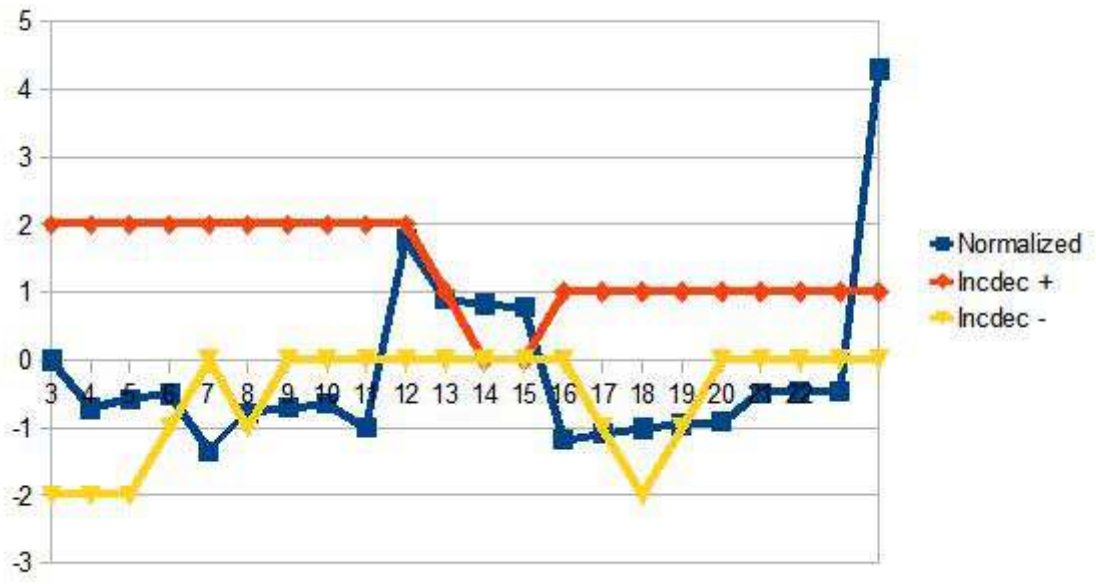

Fig. 16. Constant Step Technique Results on Outer Court Blog.

misclassified.

My third method takes a potentially more aggressive approach, that again could be informed by feedback. In this case, instead of steps in set increments, we expand or contract the limits so that the most recent false positive is just inside the limits and a false negative is just outside the limits. This ensures that a similar value with a similar classification in subsequent caches is not misclassified, however, if a misclassified cache has an error contained in normally correctly classified caches, the technique would then misclassify results it would otherwise classify correctly. To provide an example of how the limits technique behaves, please see Figure 17.

While the limits technique still maintained a decent accuracy at 75.30%, this accuracy is worse than the static method. The results for F-Score did improve over static (.255) but not as well as either the expected random results or the constant step method.

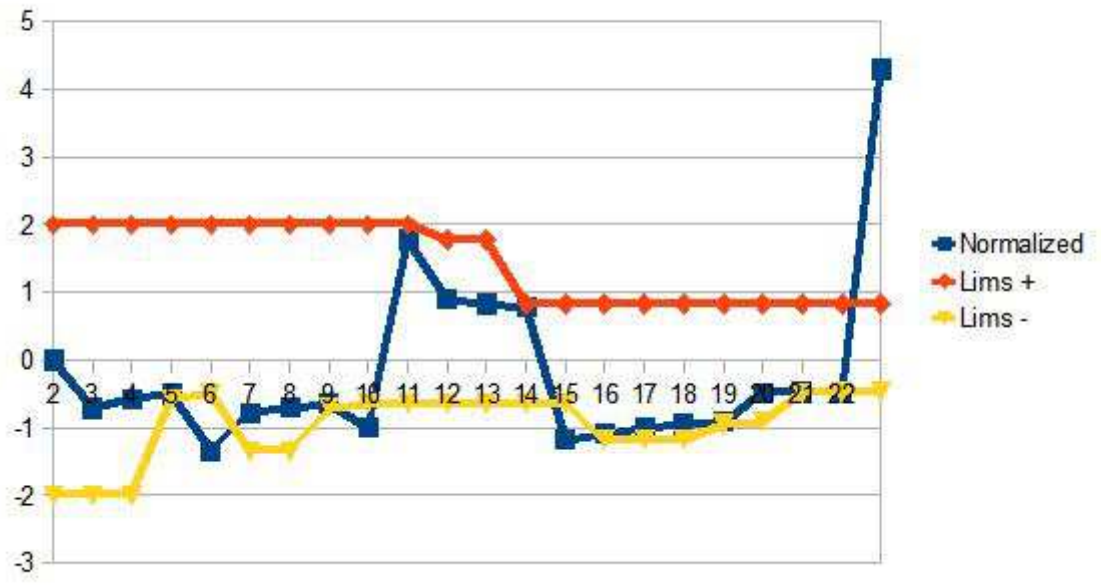

Fig. 17. Limits Technique Results on Outer Court Blog.

2. Classifier Methods

Essentially, the problem of separating normal caches and abnormal caches using the error of prediction from a Kalman Filter is a binary classification task. As such, there are a wide variety of techniques that can be used. One difference from a traditional classification task where data is split into testing and training sets, data from an online predictive method should not be simply split. Every sample becomes a new piece of testing data using the prior data as a training set. After the new point is classified and a user provides feedback on the classification, the point can be folded into the training set to generate a new classifier. I tried a variety of supervised classifier methods which produced a variety of results. Some of these results provide a trade-off of overall accuracy with increased precision and recall when compared to the heuristic methods. The classifiers utilized were variations of three categories of

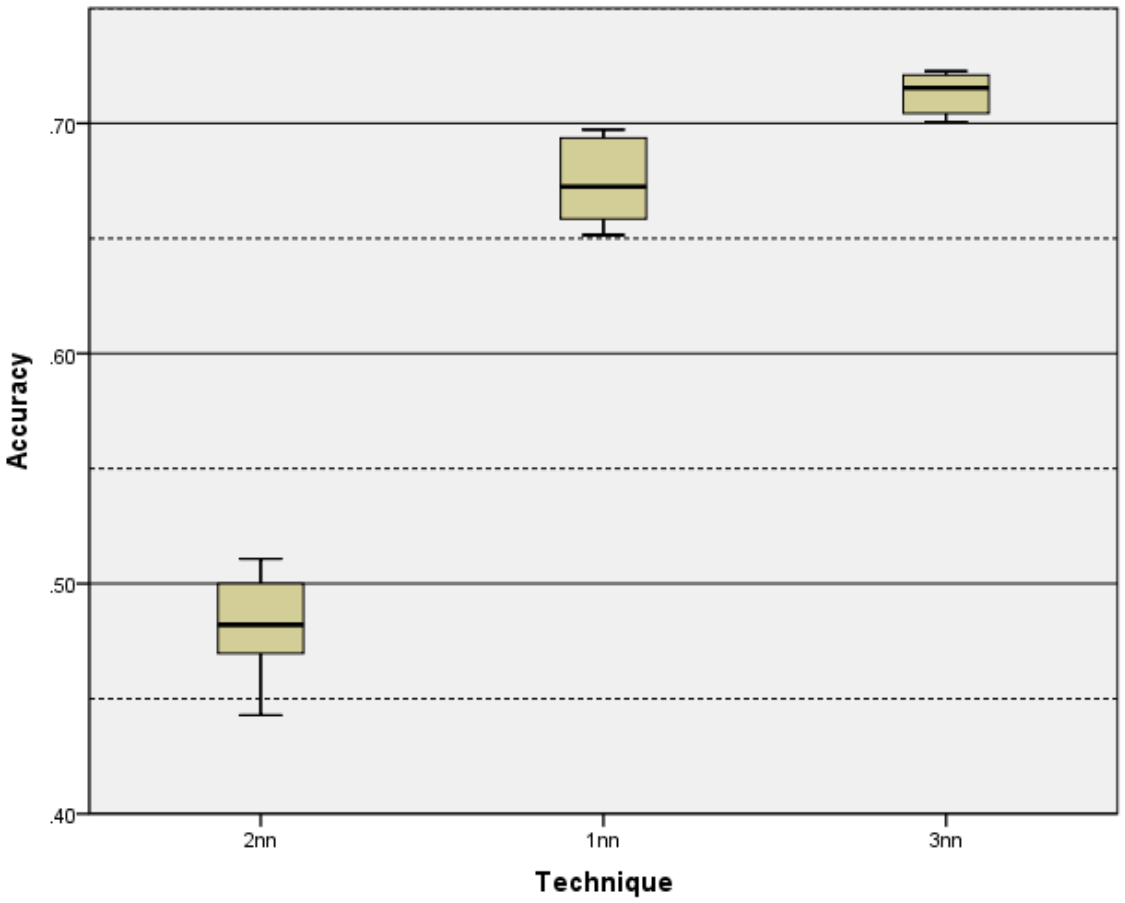

Fig. 18. Boxplot of Accuracies for kNN Methods.

binary classifiers – nearest-neighbor methods [68], support vector machine methods [69], and discriminant analysis methods [70].

Nearest-neighbor methods work by the assumption that the k nearest known samples to the test sample are likely to be of the same classification as the test sample [68]. In the case of mixed results, a majority vote of the neighbors is used to select the class. For the purposes of classifying errors, I tried 1-, 2-, and 3- nearest-neighbors variants.

The nearest neighbor approaches have mixed results as shown in Figures 18 and 19. While for accuracy all methods perform worse than the static technique, two

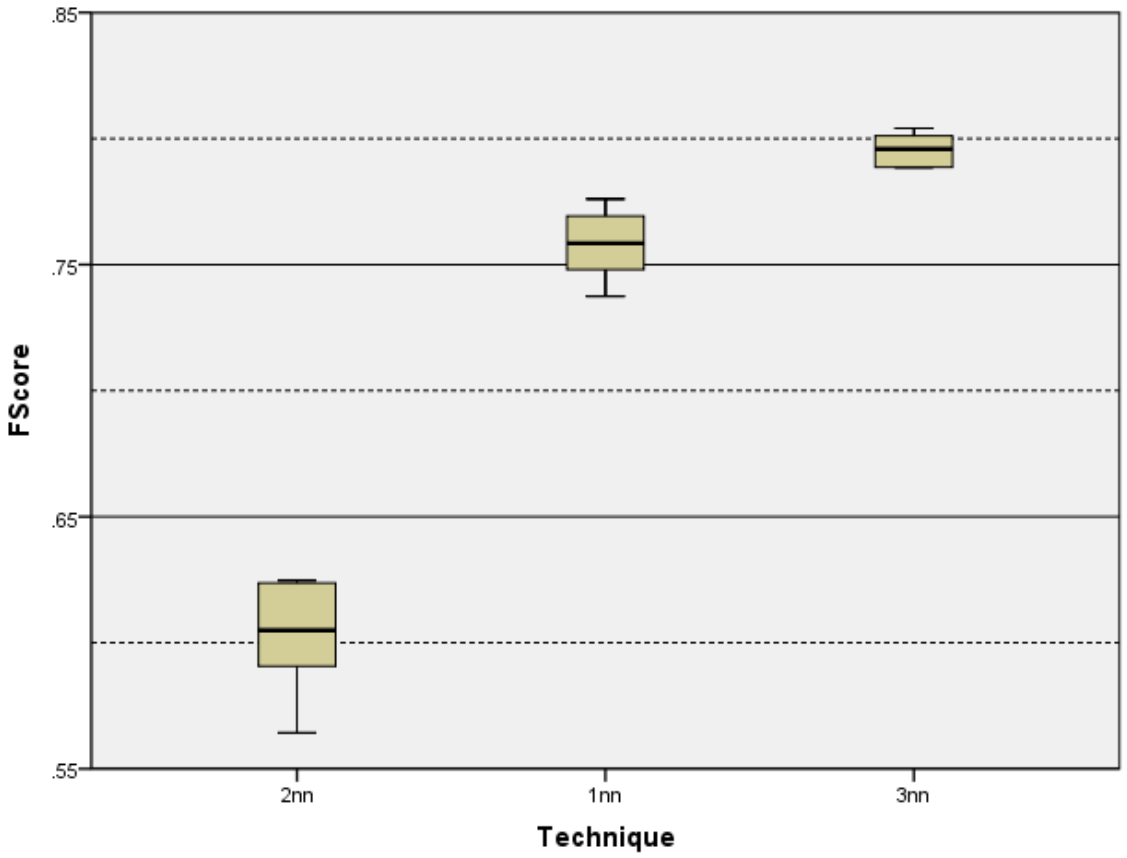

Fig. 19. Boxplot of F-Scores for kNN Methods.

methods still perform better than random chance. 3NN is the most accurate (71.35%), followed by 1NN (67.49%), and finally 2NN is worse than random at 48.19%. Conversely, by F-Score, all methods perform better than random and the static technique for all cases. 2NN is again the worse method (.602), followed again by 1NN (.758), and finally 3NN performed the best at (.796). While the constant steps technique is still the most accurate technique, all three techniques improved on the F-Score from the heuristic methods.

Support vector machines are a broad set of techniques that can be used to find a hyperplane that maximally separates two classes of data by projecting the data in

to a higher dimensionality [69]. The shape of the hyperplane is described by a kernel which can be easily replaced to provide separators of different complexities. For my tests, I tried four kernels – linear, Gaussian, terminated ramp, and the Tversky kernel. Results for these methods are presented in Figures 20 and 21.

The linear kernel is the simplest kernel in that it uses a linear hyperplane to separate the two classes [69]. While it did perform better than random for accuracy, with an accuracy of 60.56%, the linear kernel did not out-perform a static measure. While the linear kernel does provide a better F-Score than heuristic methods and 2NN at .705, 1NN and 3NN still have better F-Scores.

The next kernel in order of complexity uses a Gaussian surface to separate the two classes [69]. Gaussian does provide an improvement for both accuracy (68.76%) and F-Score (.777) over the Linear kernel. However, it did not perform statistically better than 1NN, nor did it perform statistically worse than 3NN.

Terminated ramp uses a series of linear models to improve on linear separation for oppositely labelled points that lay close to the global separating hyperplane [71]. It does this by refining the seperation with local optimal linear hyperplanes between the opposing points. Terminated ramp, while proving a much better than random accuracy (71.36%), still is less accurate than heuristic techniques. However, the F-Score for terminated ramp is the best I obtained for any method (.802). Yet, it was not statistically better than 3NN and the Gaussian kernel.

The final kernel method I used is unique from the others in that it is a non-geometric asymmetrical technique devised by Amos Tversky to match his observations of human behavior when classifying objects [72]. Despite the method being designed to better model how humans perform classification tasks, Tversky's method was the worst performing SVM method for both accuracy (69.49%) and F-Score (.599), albeit still better than random selection and 2NN for both measures in addition to having an

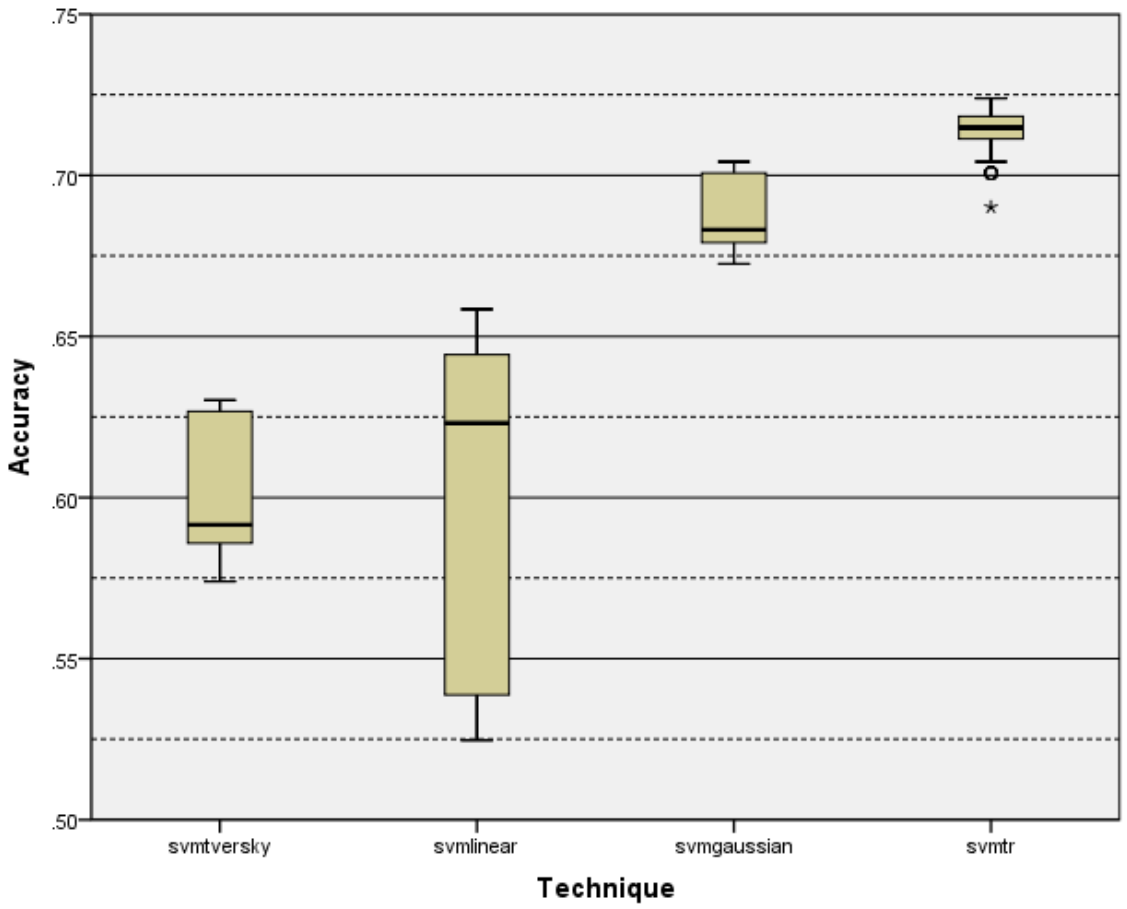

Fig. 20. Boxplot of Accuracies SVM Methods.

F-Score better than all of the heuristic methods. This is most likely due to Tversky's dependence on feature vectors with different keys. However, the preprocessing steps taken before the application of the Kalman Filter eliminate keys with a low document frequency thus homogenizing the key set.

Discriminant analysis is similar to support vector machines in that the goal is to separate two classes of data. However, discriminant analysis reduces the dimensionality of data to a single dimension with a maximal separation between the class means. While SVM methods have been shown to be more powerful, adaptable, and generally usable for binary classification tasks, they incur a complexity penalty that

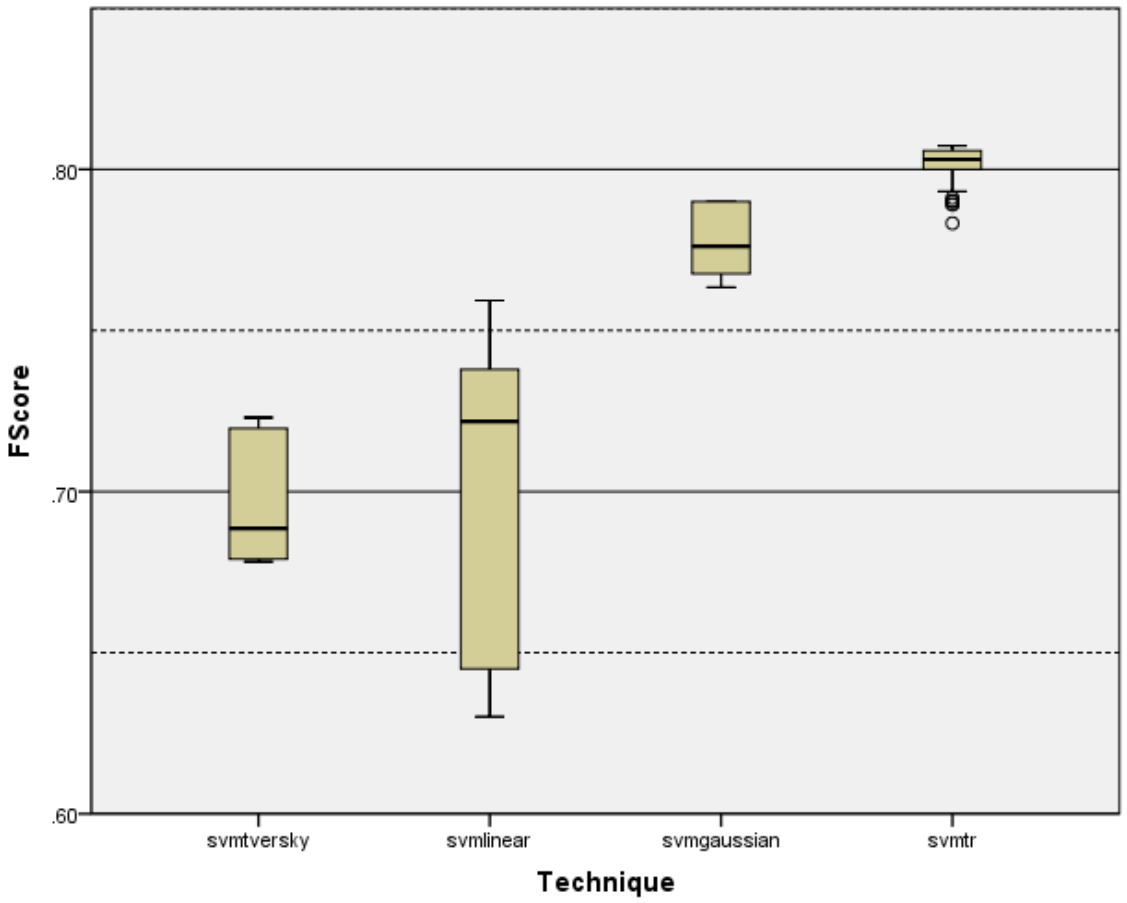

Fig. 21. Boxplot of F-Scores SVM Methods.

makes their advantages only preferable when discriminant analysis cannot provide sufficiently accurate results. For this reason I decided to evaluate the use of discriminant analysis techniques for the classification of errors. The versions of discriminant analysis I tested were Fisher Discriminant Analysis [70], Spectral Regression Discriminant Analysis [67], Penalized Discriminant Analysis [73], and Diagonal Linear Discriminant Analysis [74]. Each version uses a different discriminant function to find the maximized separator between classes. Results for the discriminant analysis methods can be seen in Figures 22 and 23.

Fisher's Discriminant (FDA) measures the ratio of the variance between classes

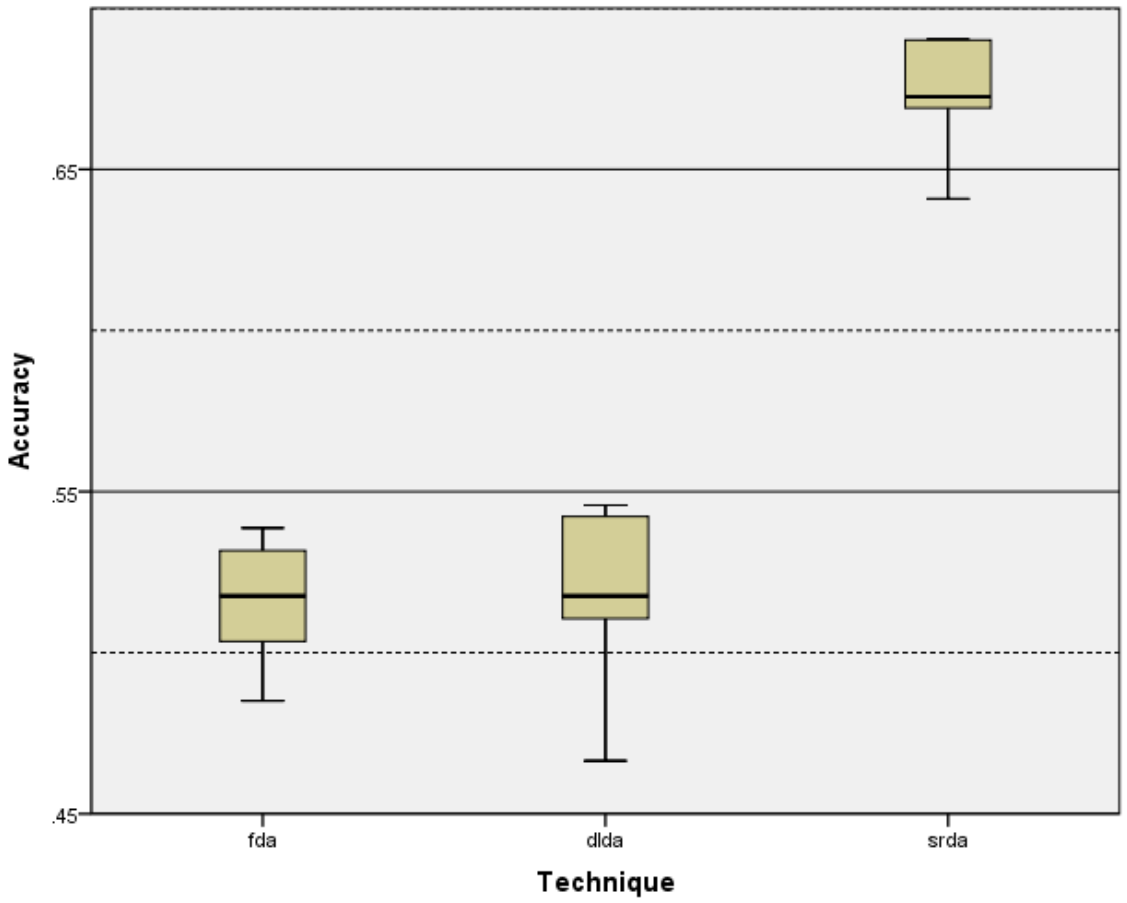

Fig. 22. Boxplot of Accuracies for Discriminant Analysis Methods.

to the variance within classes [70]. The results in terms of accuracy for FDA were better than random chance (51.60%), but not as good as heuristic methods, SVM-based methods, 1NN, or 3NN.

The second discriminant, Spectral Regression Discriminant Analysis (SRDA) [67] takes advantage of spectral analysis to reduce the complexity of the calculation of the discriminant in addition to improving overall accuracy. SRDA proved to be the best discriminant analysis technique with an accuracy of 67.41% and an F-Score of .777. While for accuracy, several methods, most notably the constant step heuristic and the static heuristic out-perform SRDA. The F-Score was statistically equivalent to

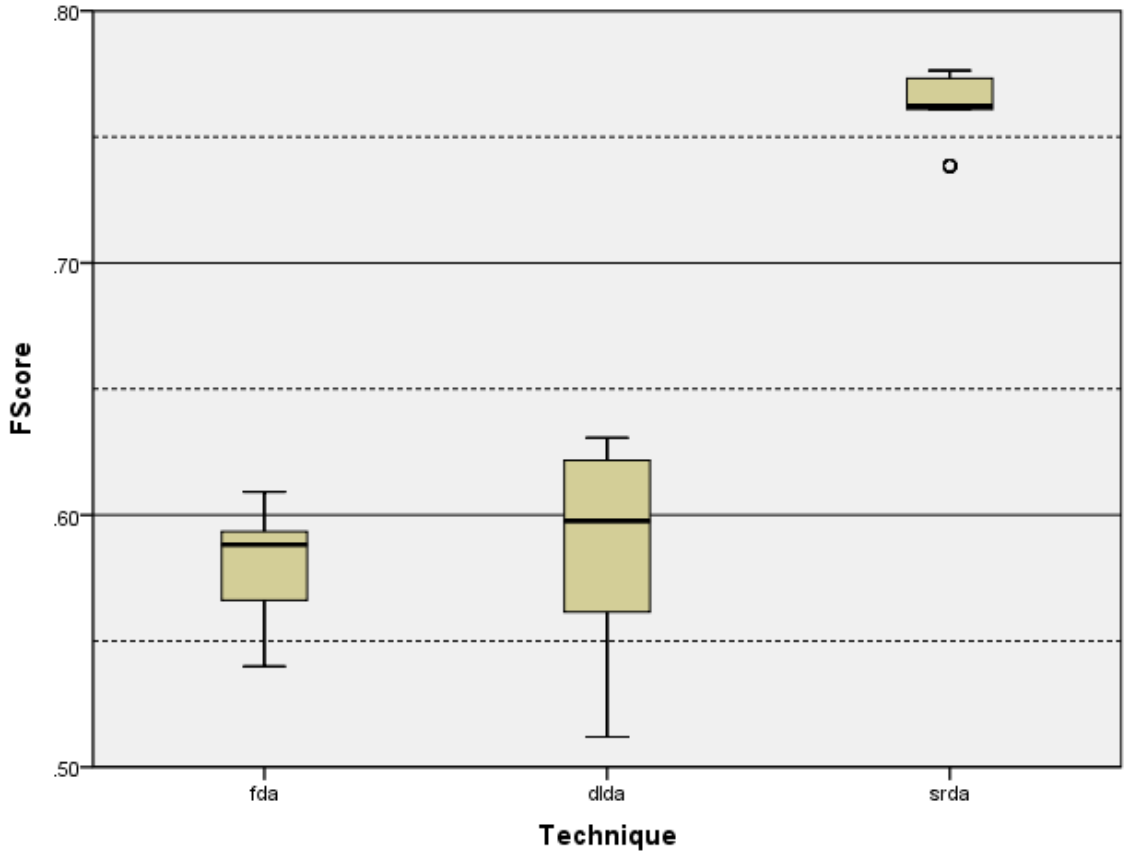

Fig. 23. Boxplot of F-Scores for Discriminant Analysis Methods.

3NN and terminated ramp, the best overall methods.

The Penalized Discriminant (PDA) [73], is a modification in Fisher's that regularizes covariance matrices to overcome deficiencies when the covariance matrix is degenerate. This can be used to overcome overfitting in large numbers of highly correlated variables or to prevent underfitting for non-linear or complex class separations. Unfortunately, PDA not only proved to be the absolute worst method, but it managed to get 0% accuracy and a 0 F-Score. Most likely this is caused by having a covariance matrix that is not degenerate.

The Diagonal Linear Discriminant Analysis (DLDA) [74], is a simplified case of

Fisher's where the covariance matrix is found to be diagonal. This allows assumptions to be made that enable a less complex algorithm to be used to find a discriminant. While DLDA proved to be more accurate (51.63%) and have a better F-Score (.588) than FDA, it was not a statistically significant improvement.

3. Statistical Outlier Detection

In previous sections I viewed the problem in terms of, first, threshold tuning, and, second, as binary classification, in this section, instead, we cast the problem as detecting statistical outliers. Outlier detection is an important area of study for mathematical statistics. Undetected outliers can skew data and in some cases can help statisticians identify a multimodal distribution in an assumed unimodal one. To address this, statisticians have sought algorithms and heuristics to identify outliers for centuries. Each method has its own drawbacks and caveats. Usually these warnings are concerning the results of overly liberal data point elimination. In our case since we are not aiming to eliminate these points, but to simply identify them for human inspection, we can proceed without fear of negative repercussions. For this class, I tested two outlier detection techniques – Chauvenet's criterion and Grubbs' test. The results for these two techniques can be seen in Figures 24 and 25.

One of the earliest techniques created was Chauvenet's criterion [75] for detecting outliers. Originally devised by observing errors in astronomical data, the method rejects values where the probability of normalized data is multiplied by the number of data points. This value is rejected as an outlier if the result is less than 0.5.

The results of my application of Chauvenet's criterion produced results that were not significantly different than the static method for both accuracy (75.75%) and F-Score (.578).

Grubbs' test [76] is a more recent method for outlier detection that was developed

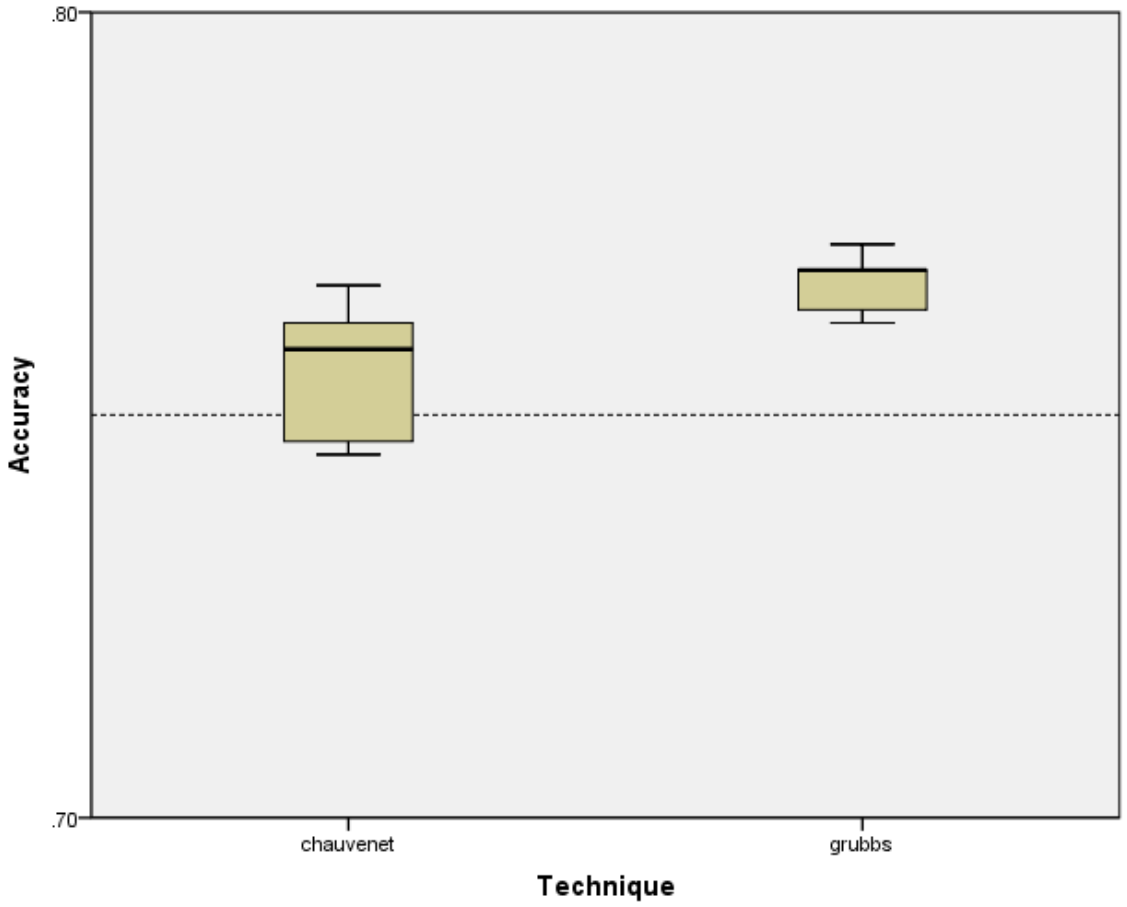

Fig. 24. Boxplot of Accuracies for Statistical Outlier Methods.

for measuring statistical accuracy for ballistic testing. The test relies on a variation of the t-statistic, defined in equation 8.5, compared to the absolute value of the standardized minimum and maximum scores in a set and runs iteratively until neither maximum nor minimum is greater than the critical region.

$$G > \frac{N-1}{\sqrt{N}}\sqrt{\frac{t^2}{N-2+t^2}} \tag{8.5}$$

While Grubbs' test performed the second best by accuracy (76.63%), it had the second worse F-Score (.010) which was not significantly any better than PDA. Additionally, Grubbs' accuracy was not significantly any better than the static heuristic.

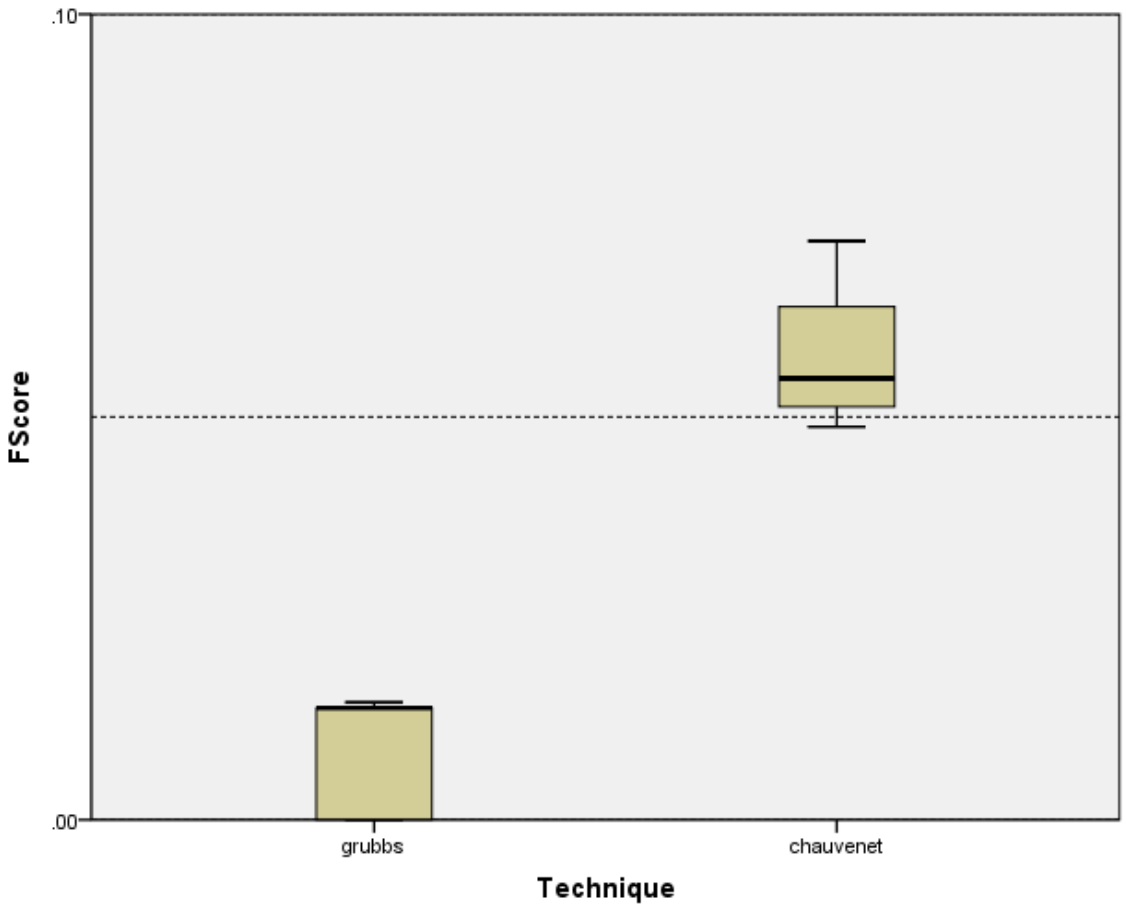

Fig. 25. Boxplot of F-Scores for Statistical Outlier Methods.

4. Ensemble Method

Finally we selected the methods with the best accuracy and F-scores and used them as votes for each cache where a majority classification ruled. The methods we used for our votes were 3NN, SVM with the Terminated Ramp kernel, SVM with the Gaussian Kernel, the constant step heuristic and Grubbs' outlier detection method. Unfortunately, as shown in Figures 26 and 27, we were not able to improve performance and only obtained a 75.32% accuracy and an F-Score of .229. I suspect that a deeper analysis of results and a weighted ensemble method could yield better results.

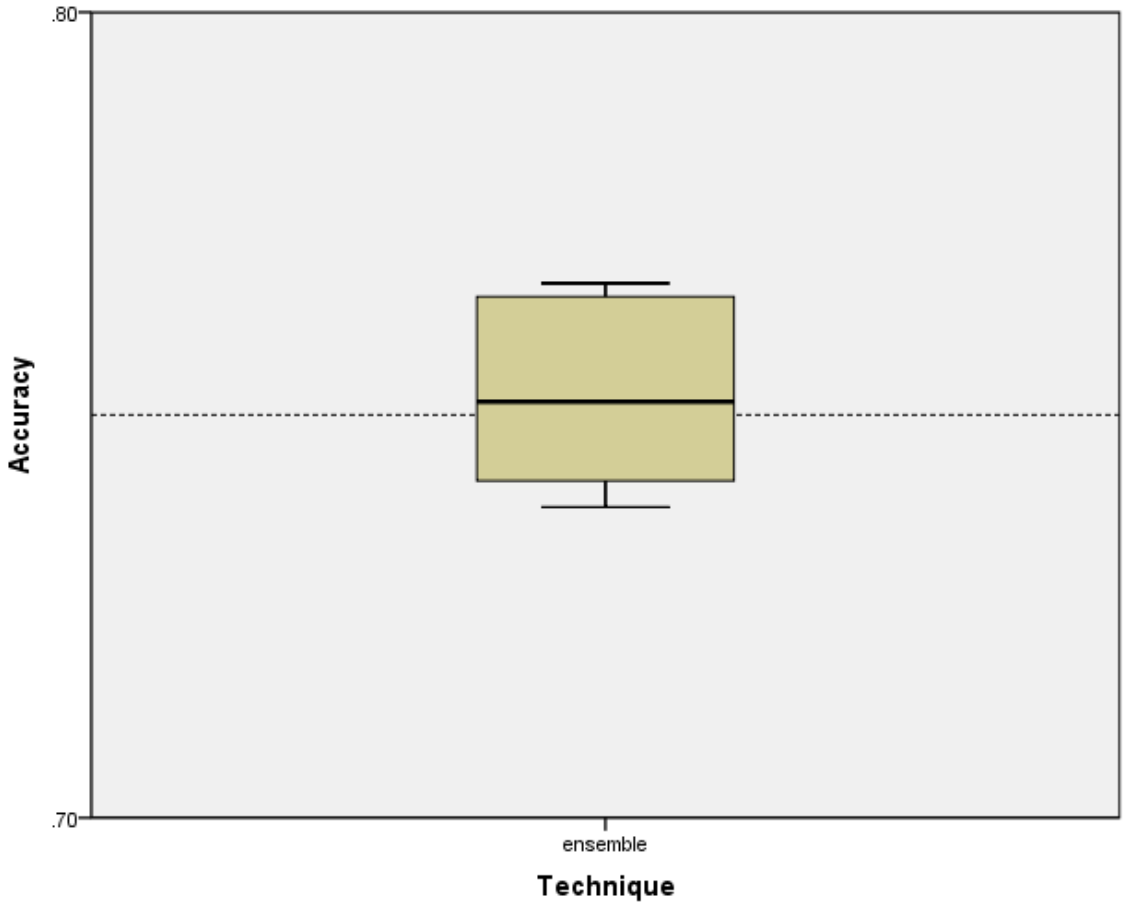

Fig. 26. Boxplot of Accuracy for Ensemble Method.

C. Kalman Filter Versus Cosine-Similarity

From the combined results of feature selection and technique selection, we identified two pairs of feature set and technique that can be objectively called the best for each of our vital methods of importance – accuracy and F-Score. For optimal accuracy we selected the static step technique using the structural change algorithm with an accuracy of 82.02%. While many other combinations of features performed equivalently to this, this single feature represented the simplest of the best accuracy techniques. For F-Score, I picked the SVM technique using the terminated ramp kernel and a fea-

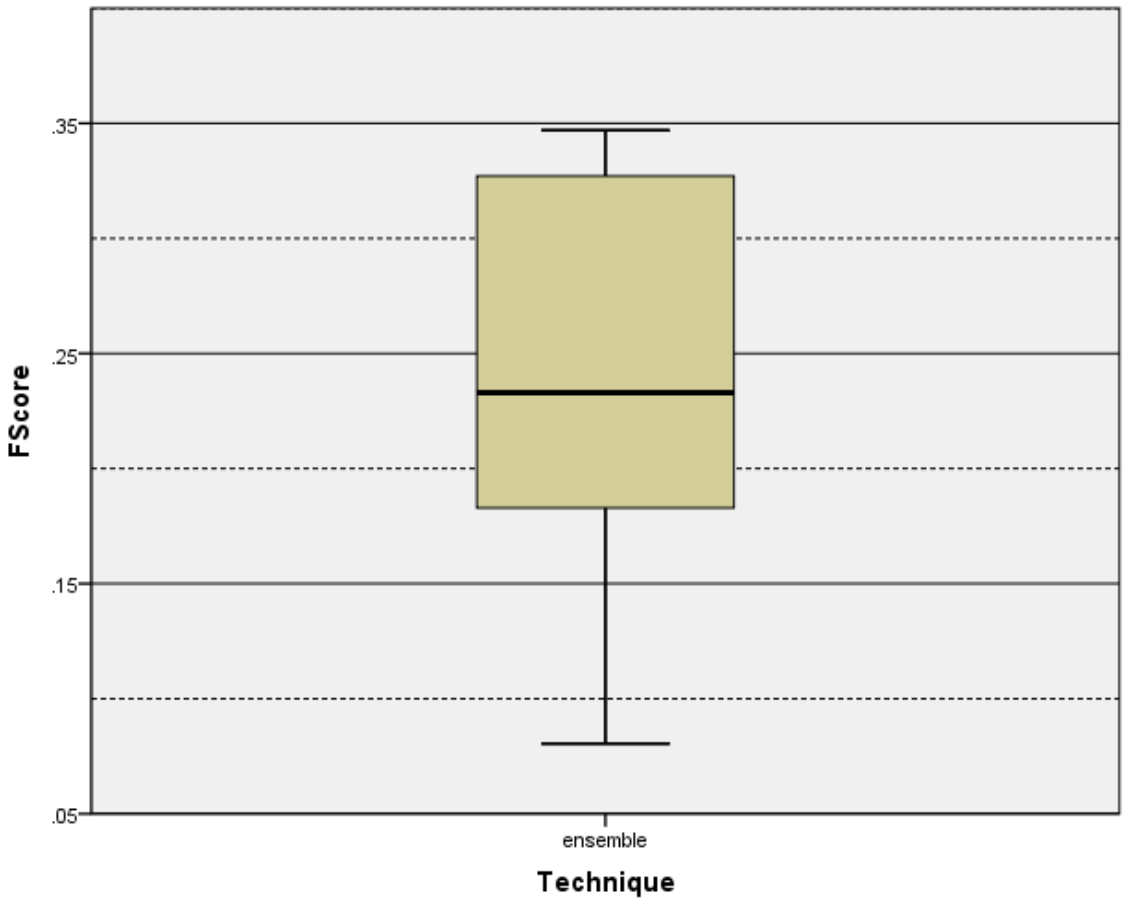

Fig. 27. Boxplot of F-Scores for Ensemble Method.

ture set consisting of structural changes and punctuation counts. This combination achieved and F-Score of .807. However, one portion of my analysis has not been objectively measured – the effectiveness of the Kalman filter technique over traditional threshold techniques. In order to address this issue, I performed cosine similarity tests both between the first and each subsequent revision and between each pair of adjacent revisions. These results are then evaluated at a variety of threshold levels to determine if the Kalman technique is better or worse than traditional techniques for every combination of feature set. Similarity is defined depending on the dimensionality of the data. If the data is one-dimensional, we use a linear distance between

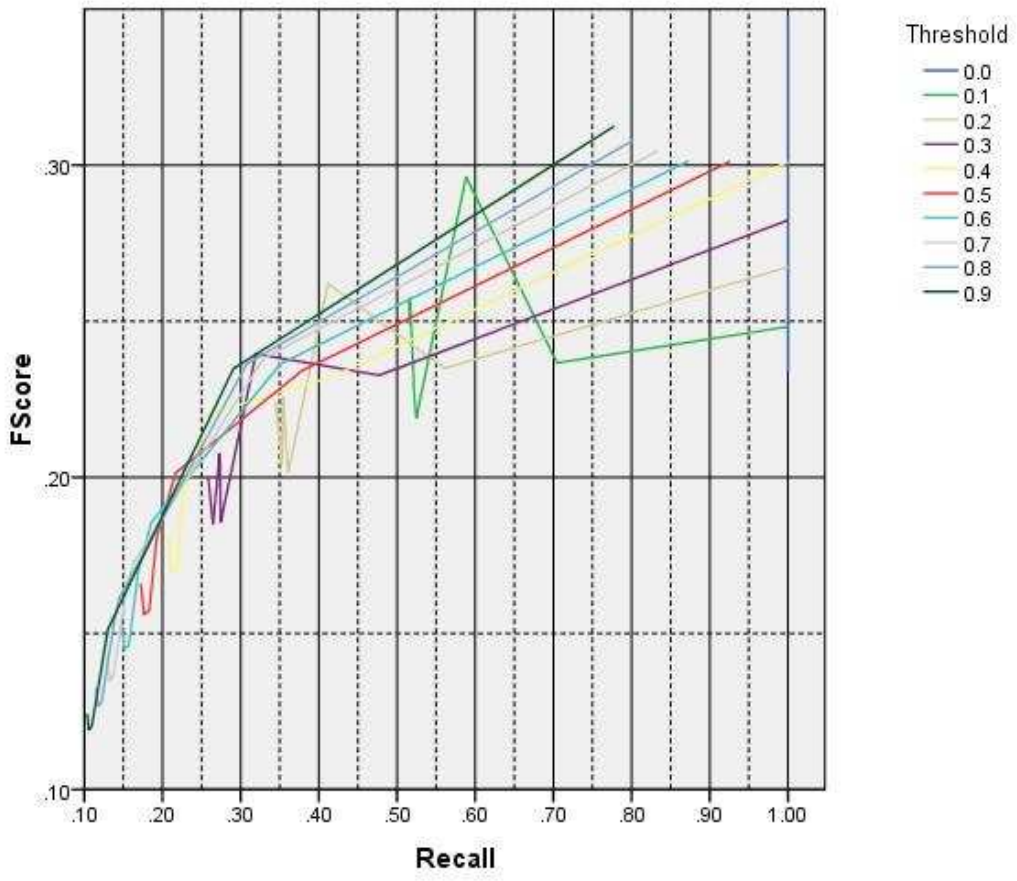

Fig. 28. F-Score and Accuracy at Multiple Relative Threshold Levels for Feature Sets.

revisions, if it is multi-dimensional, we use a standard cosine similarity metric defined for term vectors, A and B, as:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{8.6}$$

Given the usage of of both absolute and relative measures to calculate threshold, the first question is – are these different measures different and if so which one is better? Results of the threshold techniques are presented in Figures 28 and 29. I performed an ANOVA with a post-hoc Tukey's Test [77] to determine if any pair of absolute and relative measures for each threshold level were significantly different. I

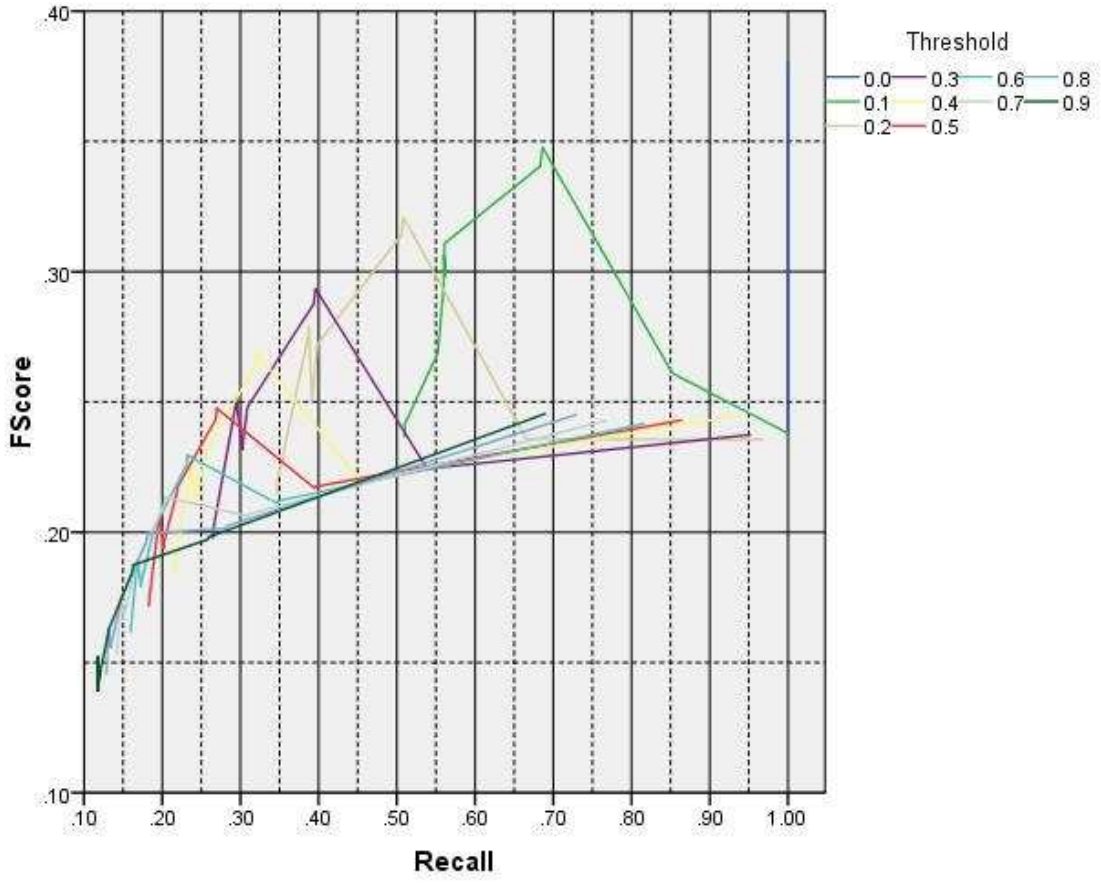

Fig. 29. F-Score and Accuracy at Multiple Absolute Threshold Levels for Feature Sets.

found no statistically significant difference between measures at any threshold level for accuracy. For the F-Scores, I found that only at the 0.1 and 0.0 threshold levels was there a significant difference between measures. In both cases the absolute measure out-performed the relative one.

I did find that different threshold levels provided an advantage over others. For accuracy, the level for the relative measure provided the best results (71.77%), while it was not statistically better than those obtained from levels at or above 0.5, it did out perform all of the levels below 0.5. For F-Score, however, the results were the opposite with the absolute method at a threshold level of 0.0 obtaining a better

F-Score (.307) than all other threshold levels.

The final step in my analysis is the comparison between threshold methods and my Kalman filter based method. As stated before, I selected the best results for each measure – SVM with a truncated ramp kernel for F-Score and the constant step heuristic for accuracy – as a point of comparison to threshold methods. In both cases, the optimal Kalman-filter based methods outperformed all of the threshold options. Just as with the comparisons between threshold options, I again used ANOVA with a post-hoc Tukey's Test to make the comparison. For accuracy, the constant step heuristic statistically outperformed the best threshold method – with a relative measure ($p = .001$). For F-Score, the SVM method with the terminated ramp kernel statistically outperformed the best threshold method – 0.0 with the absolute measure ($p = .000$). What was more surprising was that while the SVM method with the terminated ramp kernel was statistically equivalent on accuracy to the most accurate threshold method ($p = 1.000$), the constant step heuristic did statistically out perform all of the threshold tests for F-Score also ($p = .000$).

CHAPTER IX

CONTRIBUTIONS OF WORK

In summary, my contributions have been in five areas: devising a taxonomy of collections in terms of maintenance; modelling page change in terms of expectation; creating a system that allows experimentation with different features; creating a base collection for evaluating change management techniques and applying user feedback to help tune the sensitivity of change measurement.

My taxonomy of collections is focused on the relationship between locality and curatorship when maintaining a collection. This taxonomy defines four different kinds of collections that each have their own maintenance characteristics. This taxonomy holds both for digital and non-digital collections and broadens what we think of as a collection to include potential collections and collections lacking reflective purposeful intent in their assemblage.

Unlike prior work, I have embraced change as a normal and expected part of web documents. My model of a web resource is one of a changing document with an expectation of change. This means the important feature is not how much or that it changed, but the difference between the actual change and the expectation of change. This diverges from prior work, which has taken the view, similar to physical documents, that they can be treated as static artifacts that can only be changed in a limited number of ways, such as decay, damage, or adulteration.

While many changes in a physical document are readily discernible despite the loss of the document's original form, an electronic document leads to the possibility that the occurrence of change may be unknowable without the original document. This may seem to only make the view of change as detrimental stronger. However, this also illustrates that an instance of an electronic document at a particular time is not

the entire document. Instead, if change is treated as an integral part of an electronic document, a new perspective emerges that leverages the history and projected future of a document to better get a sense of what the purpose of a member of a distributed collection. While, this is an admittedly more difficult position to take, it has opened applications, such as stochastic filtering methods, that had been previously ignored.

I have also built a system that allows previously divergent measurements of change to be assembled into compound metrics to help evaluate their performance both individually and in combination with each other. This work is supported by my creation of a ground-truth corpus that allows quantitative comparisons to be made between techniques and feature sets.

Finally, by incorporating user feedback, I hope I am able to mitigate the effects of user difference in determining the importance of change by adapting analysis to each user. This helps capture the contextual differences in expectation a user has for each site in their collection.

In summary, the fundamental contribution of this work is the exploration of the essential nature of loosely controlled collections of web-based documents, and the re-evaluation of collection constituents without the limitations and assumptions that collections of physical documents impose on their users. This has lead to the creation of a set of metrics and tools that are designed around this essential nature and are thereby more useful for understanding how a web collection behaves.

The practical applications of this work are already being seen as the Distributed Collection Management Service adapts my methods for improving change detection among resources to the problem of maintaining large distributed collections of web resources currently contained in the NSDL's various Pathway project repositories.

CHAPTER X

FUTURE WORK

During the course of the creation of DCM, I made a number of decisions to restrict scope for expediency's sake. These decisions had effects on some of the overall concepts behind DCM and the design of several major systems; in particular, Decima, Phaeton, Morta, and Hannah all were constrained and limited in ways that I had not initially intended, but were necessary as the project progressed. Removing these self-imposed limits is a major future direction of DCM.

A. Searle's Chinese Room and Expectation Modelling

When conceptualizing DCM, I struck on the idea of modelling expectation instead of trying to build heuristics for dealing with sites according to the context of the collection they reside in. Expectation modelling has the advantage that I am not trying to find the absolute qualitative truth of a particular change, but instead DCM is a human-augmented system that tries to capture a user's first visceral impressions that normally determine whether they would then proceed with a deeper reading. The idea is not to definitively answer which sites have changed unexpectedly, but to draw a user away from the *prima facie* normal changes and to the cases where the status is questionable. My system has devised a solution that I have shown does a statistically better job than the existing techniques at providing results comparable to the human visceral response. However, it is not clear, and probably unlikely, that my system is undergoing the same process as a human's visceral judgement.

In the study of epistemology, Searle proposed a thought experiment in an attempt to disprove the theories of Strong Artificial Intelligence proponents known as the Chinese Room [78]. In the experiment, he supposes that a room is built with a slit in

the door through which documents written in Chinese are inserted and responses are returned [78]. Inside the room is a man who does not read or write Chinese, but has a perfect and complete book with instructions for what should be drawn given each possible set of input symbols [78]w. Searle argues that despite the correct responses being returned, there is no understanding of Chinese in the scenario since this is not how a mind processes language [78]. When this scenario is compared to my work on modelling expectation, the question arises that despite my outputs being sufficiently similar to a human's responses to a given set of input, does my system truly represent a model of expectation, and if so, can my system be said to be expecting a certain behavior? These questions depend not only on further study into the nature of expectation, but also require a deeper understanding of the epistemological theories surrounding machine intelligence and affective systems.

B. Decima 2

The current version of Decima is designed around assumptions on the characteristics of features. Namely that they can be derived from a cache independently of other caches and that the process can be expressed by a chain of Mallet filters. While this has enabled the creation of many different features, better features may depend on the history of pages. This requires a re-working of Decima from the ground in order to allow greater flexibility in feature design.

1. FAR 2

At the core of replacing Decima is a fundamentally different approach to module specification. FAR is currently based around an XML-derived configuration file that allows the construction of a Mallet filter chain. This prevents the creation of features

such as a term vector with only new words, or counts of changed structures in a structural analysis. As Decima becomes more broad in its abilities, the underlying module mechanism needs to be extended too. FAR 2 will be a full domain-specific language built using ANTLR [79]. Currently the plan is to base it around Lua [80] with extensions to support feature extraction.

C. Phaeton

Phaeton represents one of the more primitive portions of the Distributed Collection Manager. Its late creation has made it more a proof-of-concept for image change features. Representing change in images, and in fact identifying images that serve similar purposes – despite different file names – is a major area of future work for DCM. This work will require further investigation in the areas of image analysis, structural analysis, and the semantics of web pages.

D. Morta 2

Morta is an exception in the design of the Distributed Collection Manager. Unlike the other components, Morta does not currently posses modularity in its design. While Decima allows a wide variety of experimentation, Morta currently only supports Kalman filters as the back-end for analysis technique. While we did implement a cosine similarity based threshold technique, it operates outside of the Morta infrastructure. Its results are not processed by Postmorta and are not available in Hannah 2. Future plans for Morta include moving to a modular system that can be used to select from a variety of techniques which can then be compared in Hannah 3. Techniques I am considering using include: particle filters; various extensions to the Kalman filter; and other object-tracking techniques.

E. Hannah 3

Hannah 2 was designed to be an expert tool for experimentation with features for the Distributed Collection Manager. While Hannah 2 allows an expert to control every aspect of the operation of DCM, it not only does not provide an environment that is friendly to end users, but it would not be suitable for planned advancements for the other portions of DCM as described earlier in this chapter. Unfortunately, these goals are contradictory as interfaces the allow more advance experimentation would increase the system's complexity. Instead, an end-user interface should aim to provide reasonable defaults to support end-user goals. It is because of this contradiction, that I decided that future interface work for DCM would need to be split into continuing the experimentation-oriented expert tool, Hannah, and the creation of a simplified end-user system, Nyx.

In both cases one area, information visualization, would require further study and may produce novel work. For Hannah 3, information visualization work would be focused on providing techniques to allow the simultaneous comparison of results from multiple algorithms to make decisions on applicability while maintaining a recognition of drawbacks that techniques may entail. For Nyx the question is similar, but instead is focused on presenting a complex picture of a large data set in an easy to navigate interface to help evaluate changes, respond to abnormal versions, provide feedback to detection techniques, and mitigate the cost of replacing decayed resources.

CHAPTER XI

CONCLUSIONS

> It is change, continuing change, inevitable change, that is the dominant factor in society today. No sensible decision can be made any longer without taking into account not only the world as it is, but the world as it will be.
>
> Isaac Asimov - *The Encyclopedia of Science Fiction.*

To review what I have presented, first, I introduced the concept of digital distributed collections and provided examples of both their ubiquity in computing and the problematic nature of their maintenance. This was followed with a case study of different NSDL pathways and how they differ in terms of maintenance. From these examples, I constructed a taxonomy of collections in terms of how maintenance for them differs. This ties each category back to different NSDL pathways. From this taxonomy, I identified distributed collections, both curated and uncurated, as having related problems of management in terms of scale and the potential for highly dynamic behavior. To address these issues, I presented a series of problems focused around the need to identify abnormal behavior of web pages with a expectation by the collection's manager to change. I then proceeded to describe the implementation and results for a series of three preliminary studies to help illuminate the problems I had identified and to better get a sense of how to solve them. These studies informed a design for a system and an approach to analysis to help address these issues and problems. From this system, I was able to generate results on a longitudinal study of a genre of highly-dynamic web pages, blogs, which I was then able to provide quantitative results compared to a multi-judge ground-truth corpus.

My analysis of these results concluded that while different feature sets did per-

form slightly better than others, there was no significant difference and that the meaningful difference was in the method used to categorize the error in prediction from my Kalman filter approach as either expected or unexpected.

I found that the SVM method with a terminated ramp kernel provided the best F-Scores while my constant step heuristic method provided the highest overall accuracy. Since curators using a notification system are more interested in finding unexpected revisions, rather than in avoiding false-postives, we can conclude that F-Score is a more important measure than accuracy. This is especially true given the low rate of abnormal revisions in our collection. Thus, the SVM method with the terminated ramp kernel is the best overall method to use for the task of maintaining the kinds of distributed collections discussed in this research. Not only did this result hold true for techniques based on my Kalman filter approach, but also in comparison to the traditionally used threshold mechanisms.

While many areas of inquiry and experimentation do remain, as evidenced by my future work section, I am confident that my work has produced a real improvement over prior attempts to "deal with change" by embracing change and seeking out the more human notion of expectation that drives our visceral response. This artificial visceral response that I have produced serves as an initial call to attention, not an absolute definition. My work allows us to identify cases when human intervention may be needed while reducing the burden on the collection maintainer to perform the initial sifting and monitoring of highly dynamic web pages.

REFERENCES

[1] Bush, V., "As We May Think," *The Atlantic Monthly*, vol. 176, no. 1, pp. 101–108, 1945.

[2] Andreessen, M., "What's New!" http://wp.netscape.com/home/whatsnew, 1995.

[3] Yahoo Media Relations, "The History of Yahoo! - How It All Started..." http://docs.yahoo.com/info/misc/history.html, 2005.

[4] Obendorf, H. and Weinreich, H. and Herder, E. and Mayer, M., "Web Page Revisitation Revisited: Implications of a Long-Term Click-Stream Study of Browser Usage," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2007, pp. 597–606.

[5] Bogen, P. L. and Johnston, J. and Karadkar, U. P. and Furuta, R. and Shipman, F., "Application of Kalman Filters to Identify Unexpected Change in Blogs," in *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2008, pp. 305–312.

[6] Francisco-Revilla, L. and Shipman, F. and Furuta, R. and 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.

[7] Bogen P. L. and Francisco-Revilla, L. and Furuta, R. and Hubbard, T. and Karadkar, U. P. and Shipman, F., "Longitudinal Study of Changes in Blogs," in *Proceedings of the 2007 Conference on Digital Libraries*, 2007, pp. 135–136.

[8] Cassel, L. and Fox, E. and Shipman, F. and Brusilovsky, P. and Fax, W. and Garcia, D. and Hislop, G. and Furuta, R. and Delcambre, L. and Potluri, S., "Ensemble: Enriching Communities and Collections to Support Education in Computing: Poster Session," *Journal of Computing Sciences in Colleges*, vol. 25, pp. 224–226, June 2010.

[9] Shaffer, C. A. and Naps, T. L. and Rodger, S. H. and Edwards, S. H., "Building an Online Educational Community for Algorithm Visualization," in *Proceedings of the 41st ACM Technical Symposium on Computer Science Education*, 2010, pp. 475–476.

[10] Knox, D., "CITIDEL: Making Resources Available," *SIGCSE Bulletin*, vol. 34, pp. 225–225, June 2002.

[11] Tungare, M. and Yu, X. and Cameron, W. and Teng, G.F. and Pérez-Quiñones, M. A. and Cassel, L. and Fan, W. and Fox, E. A., "Towards a Syllabus Repository for Computer Science Courses," *SIGCSE Bulletin*, vol. 39, pp. 55–59, March 2007.

[12] Richards, L. M., "AMSER: Applied Math and Science Education Repository," *Reference Reviews*, vol. 24, no. 2, pp. 38–39, 2010.

[13] Ezrailson, C. M., "How Using the Physics Front Digital Library Can Support Best Practices in Science," *Meridian*, vol. 11, no. 2, 2008.

[14] Swain, D. E. and Wagy, J. and McClelland, M. and Jacobs, P., "Developing a Metadata Schema for CSERD: A Computational Science Digital Library," in *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2006, pp. 350–350.

[15] Bartolo, L. M. and Lowe, C. S. and Sadoway, D. R. and Powell, A. C. and Glotzer, S. C., "NSDL MatDL," *D-Lib Magazine*, vol. 11, no. 3, pp. 1082–9873, 2005.

[16] Plutarch, *Plutarch's Lives Complete and Unabridged in One Volume*, Hugh, A., Ed. Modern Library, 1932.

[17] Falcon, A., "Aristotle on Causality," http://plato.stanford.edu/archives/spr2011/entries/aristotle-causality/, 2011.

[18] Starr, B. and Ackerman, M. S. and Pazzani, M., "Do-I-Care: A Collaborative Web Agent," in *Conference Companion on Human Factors in Computing Systems*, 1996, pp. 273–274.

[19] Ashman, H., "Electronic Document Addressing: Dealing with Change," *ACM Computing Surveys*, vol. 32, no. 3, pp. 201–212, 2000.

[20] Furuta, R. and Shipman, F. M. and Marshall, C. C. and Brenner, D. and Hsieh, H.-W., "Hypertext Paths and the World-Wide Web: Experiences with Walden's Paths," in *Proceedings of the 8th ACM Conference on Hypertext*, 1997, pp. 167–176.

[21] Koehler, W., "Web Page Change and Persistence—a Four-Year Longitudinal Study," *Journal of the American Society for Information Science and Technology*, vol. 53, no. 2, pp. 162–171, 2002.

[22] Boese, E. S. and Howe, A. E., "Effects of Web Document Evolution on Genre Classification," in *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, 2005, pp. 632–639.

[23] Ivory, M. Y. and Megraw, R., "Evolution of Web Site Design Patterns," *ACM Transactions on Information Systems*, vol. 23, no. 4, pp. 463–497, 2005.

[24] Askehave, I and Nielsen, A. E., "What are the Characteristics of Digital Genres? - Genre Theory from a Multi-Modal Perspective," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 2005, p. 98a.

[25] Ivory, M. Y., "An Empirical Foundation for Automated Web Interface Evaluation," PhD dissertation, University of California, Berkeley, 2001.

[26] Dalal, Z. and Dash, S. and Dave, P. and Francisco-Revilla, L. and Furuta, R. and Karadkar, U. and Shipman, F., "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.

[27] Li, W. and Wu, Y. and Bufi, C. and Chang, C. K. and Agrawal, D. and Hara, Y., "PowerBookmarks: an Advanced Web Bookmark Database System and its Information Sharing and Management," in *Information Organization and Databases: Foundations of Data Organization*, Tanaka, K. and Ghandeharizadeh, S. and Kambayashi, Y., Ed. Kluwer Academic Publishers, 2000, pp. 373–385.

[28] Kellar, M. and Watters, C. and Shepherd, M., "A Field Study Characterizing Web-Based Information-Seeking Tasks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 999–1018, 2007.

[29] Abrams, D. and 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.

[30] Jones, W. and Bruce, H. and Dumais, S., "Keeping Found Things Found on the Web," in *Proceedings of the 10th International Conference on Information and Knowledge Management*, 2001, pp. 119–126.

[31] Aula, A. and Jhaveri, N. and Käki, M., "Information Search and Re-access Strategies of Experienced Web Users," in *Proceedings of the 14th International Conference on World Wide Web*, 2005, pp. 583–592.

[32] Dave, P. and Bogen, P. L. and Karadkar, U. P. and Francisco-Revilla, L. and Furuta, R. and Shipman, F., "Dynamically Growing Hypertext Collections," in *Proceedings of the 15th ACM Conference on Hypertext and Hypermedia*, 2004, pp. 171–180.

[33] Liu, H. and Ramasubramanian, V. and Sirer, E. G., "Client Behavior and Feed Characteristics of RSS, a Publish-Subscribe System for Web Micronews," in *IMC '05: Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*, 2005, pp. 3–3.

[34] Park, J. and Fukuhara, T. and Ohmukai, I. and Takeda, H. and Lee, S.-G., "Web Content Summarization using Social Bookmarks: A New Approach for Social Summarization," in *Proceedings of the 10th ACM workshop on Web Information and Data Management*, 2008, pp. 103–110.

[35] Wu, X. and Zhang, L. and Yu, Y., "Exploring Social Annotations for the Semantic Web," in *Proceedings of the 15th International Conference on World Wide Web*, 2006, pp. 417–426.

[36] Penev, A. and Wong, R. K., "Finding Similar Pages in a Social Tagging Repository," in *Proceedings of the 17th International Conference on World Wide Web*, 2008, pp. 1091–1092.

[37] Garg, S. and Gupta, T. and Carlsson, N. and Mahanti, A., "Evolution of an Online Social Aggregation Network: an Empirical Study," in *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement Conference*, 2009, pp. 315–321.

[38] Lee, K. J., "What Goes Around Comes Around: An Analysis of del.icio.us as Social Space," in *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, 2006, pp. 191–194.

[39] Lerman, K. and Galstyan, A., "Analysis of Social Voting Patterns on digg," in *Proceedings of the 1st Workshop on Online Social Networks*, 2008, pp. 7–12.

[40] Aarseth, E.J., *Cybertext: Perspectives on Ergodic Literature*. Johns Hopkins University Press, 1997.

[41] Crystal, D., *Language and the Internet*. Cambridge University Press, 2001.

[42] Henige, D., "Oral, but Oral What? The Nomenclatures of Orality and Their Implications," *Oral Tradition*, vol. 3, no. 1, p. 2, 1988.

[43] Douglis, F. and Ball, T. and Chen, Y.-F. and Koutsofios, E., "The AT&T Internet Difference Engine: Tracking and Viewing Changes on the Web," *World Wide Web*, vol. 1, no. 1, pp. 27–44, 1998.

[44] Adar, E. and Dontcheva, M. and Fogarty, J. and Weld, D. S., "Zoetrope: Interacting with the Ephemeral Web," in *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology*, 2008, pp. 239–248.

[45] Greenberg, S. and Boyle, M., "Generating Custom Notification Histories by Tracking Visual Differences Between Web Page Visits," in *Proceedings of Graphics Interface 2006*, 2006, pp. 227–234.

[46] Yadav, D. and Sharma, A. K. and Gupta, J. P., "Change Detection in Web Pages," in *Proceedings of the 10th International Conference on Information Technology*, 2007, pp. 265–270.

[47] Juola, P., "Authorship Attribution," *Foundations and Trends in Information Retrieval*, vol. 1, no. 3, pp. 233–334, 2006.

[48] Lex, E. and Juffinger, A. and Granitzer, M., "A Comparison of Stylometric and Lexical Features for Web Genre Classification and Emotion Classification in Blogs," in *Proceedings of the 2010 Workshops on Database and Expert Systems Applications*, 2010, pp. 10–14.

[49] Dowd, G. and Stevenson, L. and Strong, J., Ed., *Genre Matters: Essays in Theory and Criticism*. Intellect, 2006.

[50] Freedman, A. and Medway, P., Ed., *Genre and the New Rhetoric*. Taylor & Francis, 1994.

[51] Swales, J. M., *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990.

[52] Krippendorff, K., "Content Analysis: An Introduction to Its Methodology," *Journal of the American Statistical Association*, vol. 79, no. 385, p. 240, 2004. [Online]. Available: http://www.jstor.org/stable/2288384?origin=crossref

[53] Goldstein, J. and Sabin, R. E., "Using Speech Acts to Categorize Email and Identify Email Genres," *Proceedings of the 39th Annual Hawaii International Conference on System Sciences*, vol. 3, p. 50b, 2006.

[54] Jazwinski, A. H., *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

[55] Welch, G. and Bishop, G., "An Introduction to the Kalman Filter," *University of North Carolina at Chapel Hill*, vol. 7, no. 1, 2006.

[56] Krause, A. and Leskovec, J. and Guestrin, C., "Data Association for Topic Intensity Tracking," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 497–504.

[57] Wang, X. and McCallum, A., "Topics Over Time: A Non-Markov Continuous-Time Model of Topical Trends," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006, pp. 424–433.

[58] Cselle, G. and Albrecht, K. and Wattenhofer, R., "BuzzTrack: Topic Detection and Tracking in Email," in *Proceedings of the 12th International Conference on Intelligent User Interfaces*, 2007, pp. 190–197.

[59] Blei, D. M. and Lafferty, J. D., "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning*. New York, NY, USA: ACM Press, 2006, pp. 113–120.

[60] Van Durme, B. and Huang, Y. and Kupść, A. and Nyberg, E., "Towards Light Semantic Processing for Question Answering," in *Proceedings of the HLT-NAACL 2003 Workshop on Text Meaning*, 2003, pp. 54–61.

[61] McCallum, A. K., "Mallet: A Machine Learning for Language Toolkit," http://mallet.cs.umass.edu/, 2002.

[62] Porter, M., "Snowball Stemmers and Resources Page," On line http://www.snowball.tartarus.org. [Visited 8/22/2011].

[63] Kincaid, J. and Fishburne, R. and Rodgers, R. and Chissom, B., "Derivation of New Readability Formulas for Navy Enlisted Personnel," National Technical Information Service, Technical Report 8-75, 1975.

[64] Weaver, W., "Recent Contributions to the Mathematical Theory of Communication," *The Mathematical Theory of Communication*, vol. 1, 1949.

[65] Eaton, J. W., *GNU Octave Manual*. Network Theory Limited, 2002.

[66] Bishop, C. M., *Pattern Recognition and Machine Learning*. Springer, 2006, vol. 4.

[67] Cai, D. and He, X. and Han, J., "SRDA: An Efficient Algorithm for Large-Scale Discriminant Analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, pp. 1–12, 2008.

[68] Cover, T. and Hart, P., "Nearest Neighbor Pattern Classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21 – 27, 1967.

[69] Cristianini, N. and Shawe-Taylor, J., *An Introduction to Support Vector Machines: and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

[70] Mika, S. and Smola, A. J. and Schölkopf, B., "An Improved Training Algorithm for Kernel Fisher Discriminants," in *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics*, vol. 2001, 2001, pp. 98–104.

[71] Merler, S. and Jurman, G., "Terminated Ramp-Support Vector Machines: A Nonparametric Data Dependent Kernel," *Neural Networks*, vol. 19, pp. 1597–1611, December 2006.

[72] Tversky, A., "Features of Similarity." *Psychological Review*, vol. 84, no. 4, p. 327, 1977.

[73] Ghosh, D., "Penalized Discriminant Methods for the Classification of Tumors from Gene Expression Data," *Biometrics*, vol. 59, no. 4, pp. 992–1000, 2003.

[74] Simon, R., "Supervised Analysis When the Number of Candidate Features (p) Greatly Exceeds the Number of Cases (n)," *ACM SIGKDD Explorations Newsletter*, vol. 5, pp. 31–36, December 2003.

[75] Chavuenet, W., *A Manual of Spherical and Practical Astronomy.* Lippincott, 1871.

[76] Grubbs, F. E., "Procedures for Detecting Outlying Observations in Samples," *Technometrics*, vol. 11, no. 1, pp. 1–21, 1969.

[77] Tukey, J. W., *Exploratory Data Analysis*, ser. Behavioral Science. Addison-Wesley, 1977, vol. xvi.

[78] Searle, J.R., "Minds, Brains, and Programs," *Behavioral and Brain Sciences*, vol. 3, no. 03, pp. 417–424, 1980.

[79] Parr, T., *The Definitive ANTLR Reference: Building Domain-Specific Languages.* Pragmatic Bookshelf, 2007.

[80] Ierusalimschy, R. and de Figueiredo, L. H. and Filho, W. C., "Lua - An Extensible Extension Language," *Software: Practice and Experience*, vol. 26, no. 6, pp. 635–652, 1996.

APPENDIX A

COLLECTION USAGE SURVEY

Introduction

The purpose of this page is to provide you (as a prospective research study participant) information that may affect your decision as to whether or not to participate in this research.

You have been asked to participate in a research study to help investigate ways in which people use collections of web pages. The purpose of this study is to discover what common practices is regarding web collections in order to determine what a user needs to manage their existing web collections. Questions will be asked about your current practices and problems you are currently facing. You were selected to be a possible participant because of your potential usage of collections of web pages.

What will I be asked to do?

If you agree to participate in this study, you will be asked to answer a number of multiple choice, and free response questions. This study will take approximately 15 minutes to complete.

What are the risks involved in this study?

The risks associated with this study are minimal, and are not greater than risks ordinarily encountered in daily life.

What are the possible benefits of this study?

Your information will be used to better define user needs regarding management of collections of web pages. This will help shape ongoing work on the Distributed Collection Manager which may be of potential use to you in the future.

Do I have to participate?

No. Your participation is voluntary. You may decide not to participate or to withdraw at any time without your current or future relations with Texas A&M University being affected.

Who will know about my participation in this research study?

This study is confidential and the records of this study will be kept private. No identifiers linking you to this study will be included in any sort of report that might be published. Research records will be stored securely and only Paul Logasa Bogen II, and Dr. Richard Furuta will have access to the records.

Whom do I contact with questions about the research?

If you have questions regarding this study, you may contact Paul Logasa Bogen II at 979-862-3217 or plb@tamu.edu

Whom do I contact about my rights as a research participant?

This research study has been reviewed by the Human Subjects' Protection Program and/or the Institutional Review Board at Texas A&M University. For research-related problems or questions regarding your rights as a research participant, you can contact these offices at (979)458-4067 or irb@tamu.edu.

Participation

Please be sure you have read the above information, asked questions and received answers to your satisfaction. If you would like to be in the study, please continue.

There are 24 questions in this survey

Questionnaire

Demographic Information

A few quick questions about who you are.

What role(s) best describe you?

Please circle **all** that apply:

- Undergraduate Student
- Graduate Student
- Teacher (Primary/Secondary Education)
- Researcher
- Post-Doctorate
- Faculty (College/University)
- Alumni
- Other:

What is your age (in years)?

What is your gender?

What is your academic discipline (major)?

What is your country of origin?

Personal Collections

This section will focus on the types and purposes of web-based collections you create, use, and maintain.

Do you have a collection (or collections) of web pages that you maintain?

A collection can be something as simple as browser bookmarks, a del.ico.us account or just a list of websites.

What kind of tools do you use to store and/or maintain that collection? (Browser bookmarks, del.ico.us, Mozilla Weave)

Who uses the collection(s)? (Self, Friends/Family, Small Group, Wide Audience)

If you have multiple collections feel free to specify if they have different targeted users.

Do you use any of your collections as part of a job or for academic purposes?

How important are your collections to you?

Very low	Low	Fairly Low	Moderate	Fairly high	High	Very high

Web Page Change Tracking

This section will focus on web pages that you are interested in monitoring and the kind of changes you are interested in on these pages.

How often are there sites that you want to check for updates?

Daily	Weekly or more	2-3 times a month	Monthly	Every 2-3 months	2-3 times a year	Yearly	Rarely \Never

Please describe the kind of sites you want to check.

What kind of changes are you interested in? (Visual, Content, Interactive, etc)

Please choose the appropriate response for each item:

	Very low	Low	Fairly Low	Moderate	Fairly high	High	Very high
How dramatic are these changes?							
How difficult is it to determine if a change is important?							
How time-consuming is it to determine if a change is important?							

How often would you say that you lose track of sites in your collection?

Daily	Weekly or more	2-3 times a month	Monthly	Every 2-3 months	2-3 times a year	Yearly	Rarely \Never

Subscription Technologies.

This sections focuses on subscription technologies such as RSS.

Do you use subscription technologies such as RSS?

Please describe what you like about subscription technologies

Please describe what you don't like about subscription technologies.

System Features

This section is focused on potential features for the Distributed Collection Manager.

What capabilities should a system for maintaining collections of web pages have?

Note: The focus here is not on viewing your collections, or creating the collections, but on activities you do when you are revisiting the collection to revalidate it or check if there are updates.

How likely would you be to use a system for maintaining collections of web pages?

Very low	Low	Fairly Low	Moderate	Fairly high	High	Very high

Future Participation

This section is focused on identifying interest in further participation in DCM-related studies.

Would you be interested in participating in future studies conducted by the Distributed Collection Manager project?

Would you be interested in participating in future beta testing of the Distributed Collection Manager?

Distributed Collection Manager is a cross-browser compliant web-application host on servers at the Center for the Study of Digital Libraries at Texas A&M University.

Name?

Email Address?

Thank you for completing this survey.

VITA

Paul Logasa Bogen II was born in a blizzard in San Antonio, TX. He attended Texas A&M University and in December 2004, obtained his Bachelor of Science in computer science with a minor in philosophy. During his undergraduate, he studied abroad in Poland, interned with Intel, was President of the local ACM/IEEE-CS chapter, served on the Student Engineers' Council, and was on the Undergraduate Curriculum Committee for Computer Science. He worked as a database manager for Civil Engineering and a research assistant in the Center for the Study of Digital Libraries. Dr. Bogen obtained his Ph.D. in computer science with a certificate in Digital Humanities in December 2011. In graduate school, he worked as a Teaching Assistant, Research Assistant, and Sr. System Administrator. Additionally, he worked as a Programmer Analyst for Knowledge-Based Systems Incorporated. He has served as a mentor to both masters and undergraduate students. He has been Vice President of the Computer Science Graduate Students' Association and the Treasurer of the local chapter of Upsilon Pi Epsilon. He worked Walden's Paths, the Ensemble Project, DCM, and Ember, published articles and posters in conferences on digital libraries and computer-supported education, selected for the 2010 Virginia Tech Future Faculty Program. He currently the co-editor of the IEEE Technical Committee on Digital Libraries' newsletter. Dr. Bogen has accepted a postdoctoral position from Oak Ridge National Laboratory and can be contacted at plbogen@gmail.com or via mail at:

Applied Software Engineering Research

P.O. Box 2008, MS 6085

Oak Ridge National Laboratory

Oak Ridge, TN 37831-6085.