Computer Science Theses & Dissertations   Computer Science

Summer 2015

Detecting, Modeling, and Predicting User Temporal Intention

Hany M. SalahEldeen
Old Dominion University

DETECTING, MODELING, AND PREDICTING

USER TEMPORAL INTENTION

IN

SOCIAL MEDIA

by

Hany M. SalahEldeen
B.S. July 2008, Alexandria University, Egypt
M.S. August 2009, Universitat Autònoma de Barcelona, Spain

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2015

Approved by:

_____

Michael L. Nelson (Director)

_____

Michele C. Weigle (Member)

_____

Hussein M. Abdel-Wahab (Member)

_____

M'Hammed Abdous (Member)

ABSTRACT

DETECTING, MODELING, AND PREDICTING
USER TEMPORAL INTENTION
IN
SOCIAL MEDIA

Hany M. SalahEldeen
Old Dominion University, 2015
Director: Dr. Michael L. Nelson

The content of social media has grown exponentially in the recent years and its role has evolved from narrating life events to actually shaping them. Unfortunately, content posted and shared in social networks is vulnerable and prone to loss or change, rendering the context associated with it (a tweet, post, status, or others) meaningless. There is an inherent value in maintaining the consistency of such social records as in some cases they take over the task of being the first draft of history as collections of these social posts narrate the pulse of the street during historic events, protest, riots, elections, war, disasters, and others as shown in this work.

The user sharing the resource has an implicit temporal intent: either the state of the resource at the time of sharing, or the current state of the resource at the time of the reader "clicking". In this research, we propose a model to detect and predict the user's temporal intention of the author upon sharing content in the social network and of the reader upon resolving this content. To build this model, we first examine the three aspects of the problem: the resource, time, and the user.

For the resource we start by analyzing the content on the live web and its persistence. We noticed that a portion of the resources shared in social media disappear, and with further analysis we unraveled a relationship between this disappearance and time. We lose around 11% of the resources after one year of sharing and a steady 7% every following year. With this, we turn to the public archives and our analysis reveals that not all posted resources are archived and even they were an average 8% per year disappears from the archives and in some cases the archived content is heavily damaged. These observations prove that in regards to archives resources are not well-enough populated to consistently and reliably reconstruct the

missing resource as it existed at the time of sharing. To analyze the concept of time we devised several experiments to estimate the creation date of the shared resources. We developed Carbon Date, a tool which successfully estimated the correct creation dates for 76% of the test sets. Since the resources' creation we wanted to measure if and how they change with time. We conducted a longitudinal study on a dataset of very recently-published tweet-resource pairs and recording observations hourly. We found that after just one hour, ∼4% of the resources have changed by ≥30% while after a day the change rate slowed to be ∼12% of the resources changed by ≥40%.

In regards to the third and final component of the problem we conducted user-behavioral analysis experiments and built a dataset of 1,124 instances manually assigned by test subjects. Temporal intention proved to be a difficult concept for average users to understand. We developed our Temporal Intention Relevancy Model (TIRM) to transform the highly subjective temporal intention problem into the more easily understood idea of relevancy between a tweet and the resource it links to, and change of the resource through time. On our collected dataset TIRM produced a significant 90.27% success rate. Furthermore, we extended TIRM and used it to build a time-based model to predict temporal intention change or steadiness at the time of posting with 77% accuracy. We built a service API around this model to provide predictions and a few prototypes. Future tools could implement TIRM to assist users in pushing copies of shared resources into public web archives to ensure the integrity of the historical record. Additional tools could be used to assist the mining of the existing social media corpus by derefrencing the intended version of the shared resource based on the intention strength and the time between the tweeting and mining.

Dedicated to the ones I love.

بسم الله الرحمن الرحيم

ACKNOWLEDGMENTS

This is to all the ones who supported me, loved me, and pushed me through. Without you none of this would be possible. A PhD is a marathon and you have been with me in every step of the way. In moments of triumph, and especially in the moments of doubt, frustration, and despair. For that, I can't thank you enough, but in here I will try.

First of all, I thank God for everything, and I do it in every prayer. Not only for enabling me to reach the end of this feat, but also for providing the environment for me to make it possible, with a loving family, and amazing friends. In the Quran God said "You haven't received of knowledge except a small amount", the more I acquire of knowledge the more humbled I get.

I was lucky enough to be raised in a loving home with my mother Laila and my father Salah who I owe everything I am and will ever be. I thank them for believing in me when I myself had doubts. For keeping me strong, and supporting me till this day with their love and understanding. I thank my sister Noha who is the princess of my heart and the one who I spent the best childhood with.

Words cannot express my thanks and appreciation towards my advisor Dr. Michael L. Nelson not only for guiding me through the rough seas that is the PhD, but also for being there for me and mentoring me in research and in life. He not only advised my research progress, but was always there to give me life advise, to calm me down at times, and to encouraged me in others, and was my anchor in the moments of despair (which was a lot!). He also made me love American football and restoring classic automobiles too!

I especially thank my great co-advisor Dr. Michele C. Weigle for her understanding, support, and guidance, not to mention being really patient with me. I thank Dr. Hussein Abdel Wahab for being the father figure here for me in the department and helping me since the beginning of my term here in the department. I thank Dr. Abdous for his guidance and support in research and at the islamic center, and always being there for me and my brothers in the community and providing us with invaluable life advice.

I thank my roommate, and the brother I never had, Wassim Obeid for his support, awesome conversations, and advice. Not to mention helping me in my hobby

ix

Herman, for providing me with a warm environment to work on my research and filling me with great coffee. Half of this dissertation has been written there while sipping their amazing java. Essam Nazef who has been and always will be my older brother and someone I can rely on and seek his advice. I thank Aunt Mona, Marwan and Hassan ElRakabawy for being my warm supportive family here in the US. I also thank Mrs Janet Brunelle for her continuous support.

Last and most definitely not least I thank Tarek Abdou, Hussein ElZawawy, Adham ElNawawy, Mohamed ElDesouky, Emad Elwany, and all my family and friends here in the US and back home. I am the sum of all your care and love. Without you all, none of this would have been possible. I sincerely apologize if I didn't mention someone explicitly or implicitly. I am forever in your debt.

In search for my personal legend...

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

"A journey of a thousand li starts beneath one's feet."

— Lao Tzu, *The Tao Te Ching*

Covering a war means going to places torn by chaos, destruction, and death, and trying to bear witness. It means trying to find the truth in a sandstorm of propaganda when armies, tribes or terrorists clash. And yes, it means taking risks, not just for yourself but often for the people who work closely with you.....Many of you here must have asked yourselves, or be asking yourselves now, is it worth the cost in lives, heartbreak, loss? Can we really make a difference?......Our mission is to report these horrors of war with accuracy and without prejudice. We always have to ask ourselves whether the level of risk is worth the story. What is bravery, and what is bravado? – Marie Colvin[1]

These were excerpts from the award-winning American journalist and war correspondent Marie Colvin's speech at St. Brides Church in London in 2010 commemorating journalists and their support staff who gave their lives to report from the war zones of the 21st Century (Figure1a). In 2001, she lost her left eye in a rocket propelled grenade (RPG) explosion while covering the Sri Lankan Civil War. In less than two years after this speech, Colvin lost her life in an explosion in February 2012 while crossing into Syria on the back of a motorcycle to cover the Syrian civil war along with a colleague and war correspondent, the award-winning French photographer Rémi Ochlik (Figure1b).

Ochlik was 28 when he died shortly after arriving in Homs, Syria. A couple of months before that he was in Libya covering the fall of Tripoli. In 2011, Ochlik was at the heart of the Jasmine Revolution in Tunisia, where he was with his friend and colleague Lucas Dolega, also a French photographer, who died shortly after

---

[1]http://www.theguardian.com/commentisfree/2012/feb/22/marie-colvin-our-mission-is-to-speak-truth

being shot by the Tunisian police. Less than a month after that, Ochlik was at the Tahrir Square in Egypt near where the Egyptian journalist Ahmed Mohamed Mahmoud was shot to death by a police sniper while filming the riot police throwing tear gas canisters into the crowds of protesters during the 18 days of the Egyptian Revolution.

Two years later in 2013 and a few klicks away from where Ahmed Mahmoud had died, Ahmed Samir Assem an Egyptian freelance photographer, at the age

(a) Marie Colvin in Tahrir Square in Egypt (courtesy of theguardian.com)

(b) Photographer Rémi Ochlik (courtesy of bbc.co.uk)

Figure 1. Late war correspondents who died in Syria 2012

of 26, (Figure2a) captured his own death through the lens of his camera (Figure 2b). Ahmed was shot in the forehead by an army sniper while filming on top of the buildings during pro-Morsi protests outside the Republican Guard building in Cairo, where some believe the ousted president Mohamed Morsi was being held.

The common thread between all of these unsung heroes is that they all gave

(a) Ahmed Assem in Cairo 2013 (courtesy of quebec.huffingtonpost.ca)

(b) Ahmed Assem filming an army sniper seconds before he shot him (courtesy of nydailynews.com)

Figure 2. Late photographer Ahmed Samir Assem

their lives while trying to capture the reality on the ground during times of war, revolution, or conflict and convey this reality through their words, photographs, or films. On a larger scale during the Arab Spring, hundreds of civilians were injured, killed or mutilated while protesting, taking photographs on their cellular phones of riot police brutality, tweeting the pulse of the street second by second, or even spurring the protests and warning others of ambushes. These circumstances place a huge value in all of this content captured and published in social media narrating the incidents and giving unfiltered insights for future generations and historians to know exactly what was happening in these turning points in history. In the following years, these contemporary tweets, videos, pictures, and Facebook posts would tell what a thousand articles, written at a later time, could not convey.

1.1 INTENTION AND THE CHANGING WEB

Given that Ochlik lost his life in pursuit of journalism, one could argue that his works and the reactions of his adorers will withstand the test of time, but this is unfortunately not the case. After just one and a half years following his death the content related to him is already disappearing.

The Wikipedia page about him has eight external links about his life, three of which are missing from the live web as shown in Figure 3. Even content tweeted depicting his death started disappearing as well by deletion of the entire wordpress website, as shown in Figure4.

His girlfriend Emilie Blachère wrote a touching love letter on the first anniversary of his death mourning him, which was tweeted by hundreds of followers as shown in Figure 5. This letter also went missing from "le journal de la photographie" website after its shutdown in August 2013[2].

Ochlik's friends and colleagues curated his photographic work about the Arab Spring Revolutions and posthumously published it in 2012 as "Révolutions, du rêve au printemps de Rémi Ochlik"[3]. To add insult to injury, the website was down upon writing this document, as shown in Figure 6.

Content on the web is in constant danger of loss or deletion. This could be for various reasons, among which is deliberate deletion by authors or system administrators. This deletion could be due to limited space on servers or fear of reprisal

---

[2]http://www.lemonde.fr/culture/article/2013/08/30/le-journal-de-la-photographie-ferme-ses-portes_3469146_3246.html

[3]http://www.webullition.info/mjmn/portrait-de-remi-ochlik/
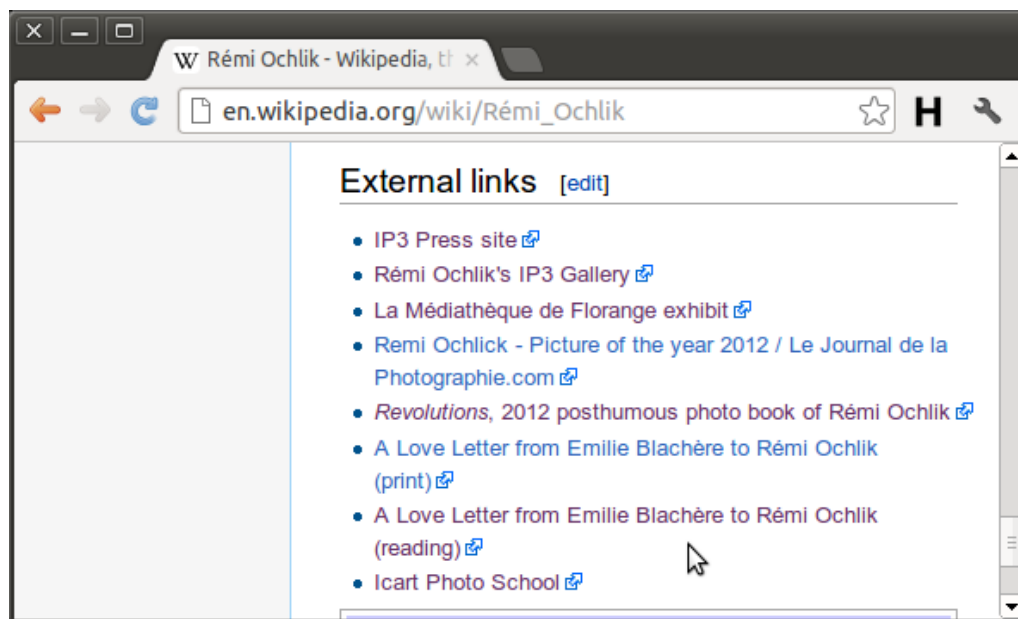
(a) 3 out of 8 Wikipedia external links entries are missing
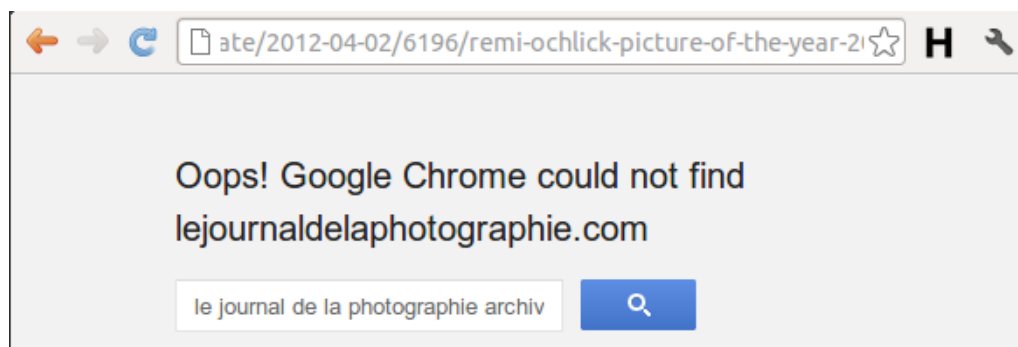
(b) Article link from Wikipedia website. `http://lejournaldelaphotographie.com/archives/by_date/2012-04-02/6196/remi-ochlick-picture-of-the-year-2012`

Figure 3. Rémi Ochlik's Wikipedia page

(a) Tweet about Rémi Ochlik's death. `https://twitter.com/gwynelora/status/278507367325368320`
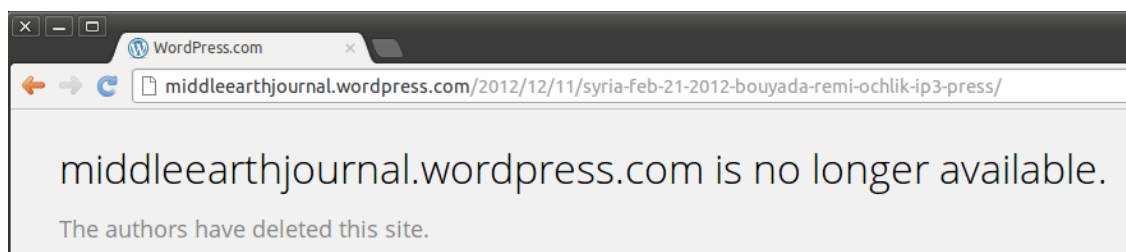
(b) Wordpress website deleted with article about Ochlik. `http://middleearthjournal.wordpress.com/2012/12/11/syria-feb-21-2012-bouyada-remi-ochlik-ip3-press/`

Figure 4. Articles missing about Rémi Ochlik's death

in the case of controversial content. During and after the Egyptian Revolution of 2011, a multitude of biased journalists and corrupt politicians who supported the ousted president Mubarak deleted their published articles from news portals shortly after the success of the revolution. This has not only happened in Egypt; the European Union court passed a ruling to "the right to be forgotten" which forced search engines like Google to remove links to certain web pages in March 2014[4]. This is

---

[4]`http://blogs.telegraph.co.uk/news/douglascarswellmp/100271108/europe-tells-`

(a) A tweet depicting the love letter. `https://twitter.com/anisandwith/` `status/329572586151346176`
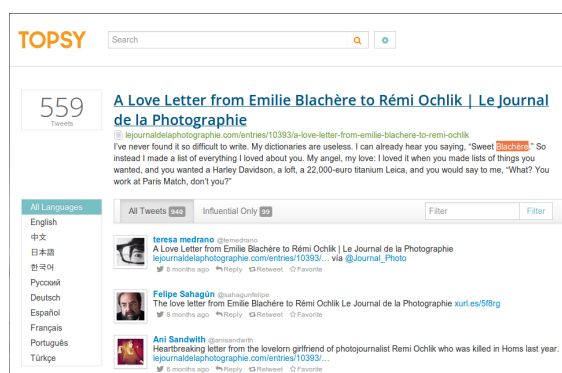
(b) Tweets linking to this letter.
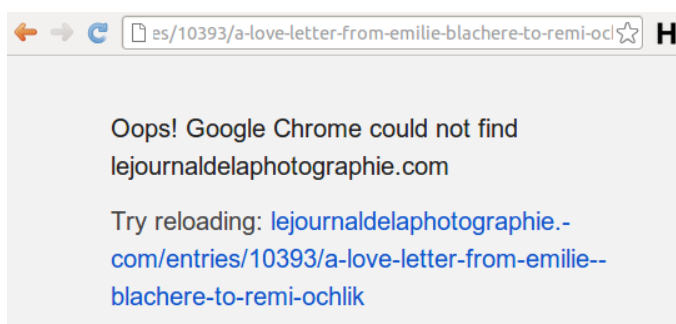
(c) The website with a copy of the love letter is shutdown. `http://` `lejournaldelaphotographie.com/entries/10393/a-love-letter-from-` `emilie-blachere-to-remi-ochlik`

Figure 5. Emilie Blachère's love letter to Rémi Ochlik

`google-to-delete-data-corrupt-politicians-will-be-thrilled-now-they-can-hide-` `their-secrets/`
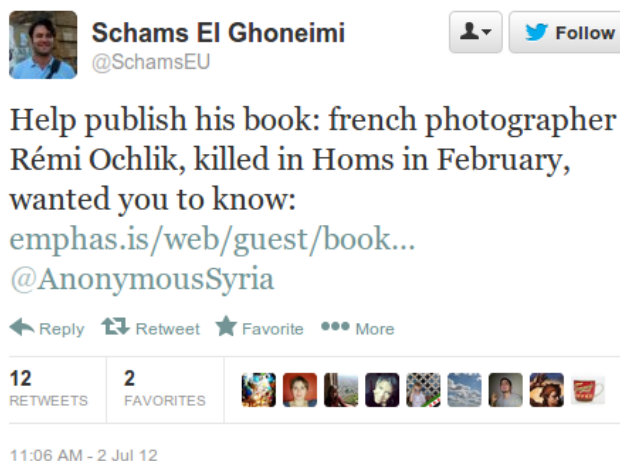
(a) A tweet advertising Ochlik's book. `https://twitter.com/SchamsEU/status/219854532212031488`

(b) The website publishing Remi's book is down. `http://www.emphas.is/web/guest/bookproject?projectID=695`

Figure 6. Rémi Ochlik's posthumous book "Révolutions"

viewed by many as corrupting history and giving a free pass to corrupt politicians and criminals to erase the past[5]. Also, services and Internet companies are prone to shutdowns all the time. Some well-known examples include the shutdown of GeoCities by Yahoo! Inc.[6] and Tr.im URL shortener[7] services both in 2009, Google Wave

---

[5] `http://cpj.org/blog/2014/06/eu-right-to-be-forgotten-ruling-will-corrupt-histo.php`

[6] `http://content.time.com/time/business/article/0,8599,1936645,00.html`

[7] `http://mashable.com/2009/08/09/trim-shuts-down/`

in 2010[8], and Google Reader in 2013[9].

Beside the danger of loss, web content faces the subtle and more pressing danger: alteration. When you read a tweet about Ochlik, click on the associated link, and find this webpage missing, you will know implicitly that this is not what the original author intended for their followers to see. But if the page was changed from the state the author saw at the moment of sharing, a bigger problem arises. This causes an inconsistency in the web and a mismatch between what the author intended for you to see and what you are actually seeing right now.

This inconsistency is sometimes negligible, for example, when it is only the change in the *timestamp* on the page. Other times it is intentional, as in the case of advertisements displayed on the page. These advertisements are intended to be different at each point in time to ensure exposure and diversity to the sponsoring companies. Also the change could be intentional if the page had a comments section and other users keep appending their comments on the intended article. In these two cases, the change is tolerable, and maybe desirable too. In other cases, the intended article could be completely replaced with something more contemporaneous. Or worse the author may alter or removes certain paragraphs to change the direction of the posted article, like what happened with NBC News who retracted and edited a controversial article about ObamaCare[10] on October 29th, 2013. These cases are less detectable and affect the consistency of the conveyed intended story dramatically.

Beside the numerous missing resources (Figures 4, 5, and 6) that are linked in tweets and posts about Ochlik and his life-work, in multiple tweets we found that the links actually direct the reader to content that is completely unrelated to Ochlik as shown in Figures 7, 8, and 9.

As the examples demonstrate, the tweet did not change and the resource referenced in it did not disappear. Instead, the resource changed, which renders the tweet and the story incoherent and inconsistent. There is an obvious mismatch in the temporal intention of the author and what is perceived by the reader. These scenarios illustrate the problem we are trying to detect and solve. We coined the term *Temporal Intention* to differentiate between what was intended by the author at the time of publishing the social post t_{tweet} and what is perceived by the reader

---

[8]http://www.cnet.com/news/google-pulls-plug-on-google-wave/

[9]http://googlereader.blogspot.com/2013/07/a-final-farewell.html

[10]http://www.ijreview.com/2013/10/90544-watch-nbc-news-drops-bombshell-obama-lying-obamacare-tries-redact-article/

(a) A tweet about Ochlik on MSNBC posted in 2012. ($t_{tweet} = 2012 - 02 - 22$)
https://twitter.com/NBCNewsPictures/status/172368454551212032
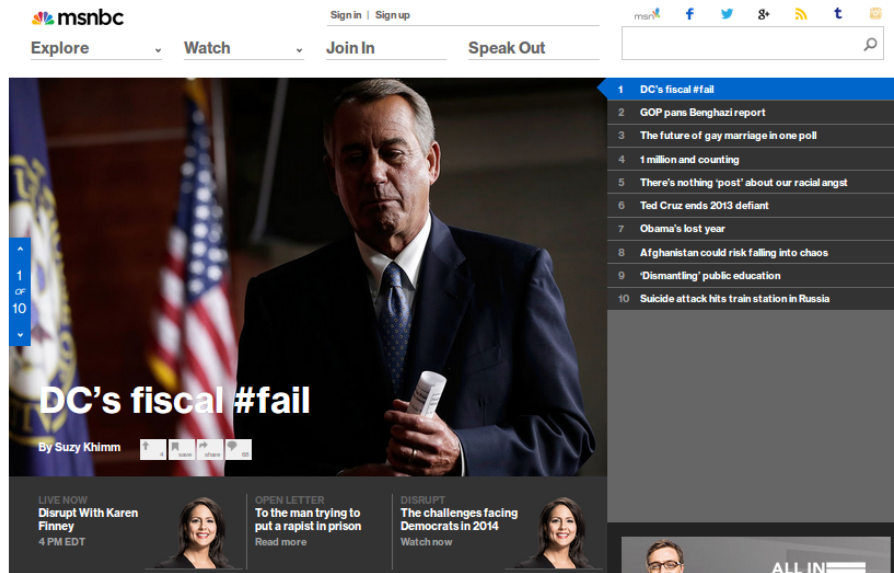
(b) The MSNBC website's current state upon clicking on the link in the tweet. ($t_{click} = 2014 - 07 - 22$) http://www.msnbc.com/

Figure 7. MSNBC tweet featuring Ochlik's work visited in 2014

at the time of reading the post t_{click}.

1.2 RESEARCH STATEMENT

The ability to share web resources is one of the key factors that makes social media universally appealing. For a variety of reasons, this sharing is done by reference

(a) A tweet about Ochlik in the British Journal of Photography. ($t_{tweet} = 02/23/2012$)
https://twitter.com/ronhaviv/status/172769102417502208
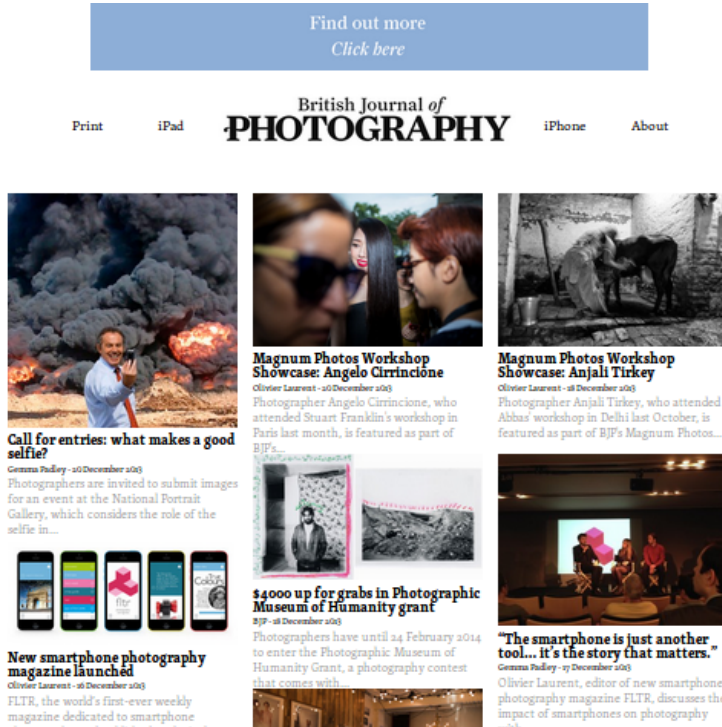
(b) The British Journal of Photography website upon clicking on the link in the tweet.
($t_{click} = 07/22/2014$) http://www.bjp-online.com/

Figure 8. British Journal of photography featuring Ochlik's work

(e.g., tweeting a URI (Uniform Resource Identifier), typically with a personalized URI alias constructed at the time of the tweet). If one shares a URI on twitter and their followers read it immediately, then there is a good chance that the state of
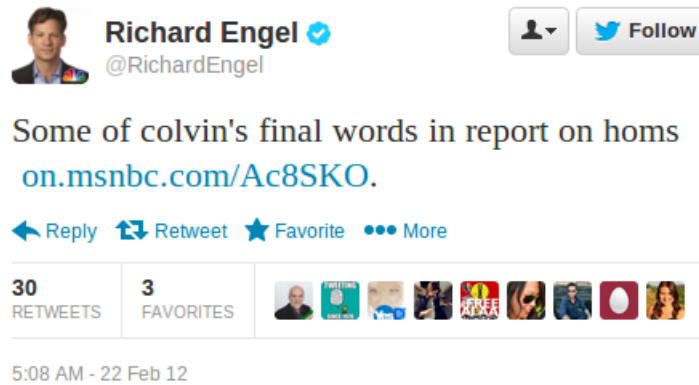
(a) A tweet about Marie Colvin's last words. ($t_{tweet} = 02/22/2012$)
https://twitter.com/RichardEngel/status/172306786643218433
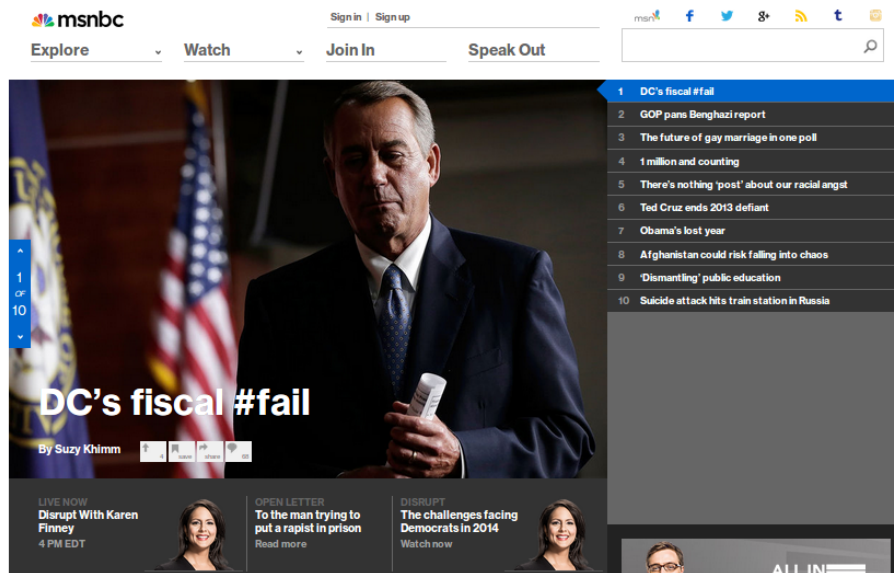
(b) The MSNBC website's current state upon clicking on the link in the tweet. ($t_{click} = 07/22/2014$) http://www.msnbc.com/

Figure 9. MSNBC tweet featuring Marie Colvin's last words

the shared resource has not changed. However, if they (re-)read later, the state of the resource has almost certainly changed. In some cases this change is desirable and not problematic as we stated before. For other resources, the changed state can introduce ambiguity and confusion. A need arises for an exploration of the concept of temporal intent in the act of sharing a URI: *did I mean to share the most current version, or the version archived at the time of share?* Although the web does not provide a direct mechanism for accessing prior states of a resource, prior states can

be accessed via web archives like the Internet Archive. However, archival coverage is uneven and few people are even aware of the existence of archives. Thus there can arise a temporal discrepancy between the resource at the time the page author created a link to it (t_{tweet}) and the time when a reader follows the link (t_{click}).

If social media is supplanting journalism as the "first rough draft of history", then we cannot assume the time between sharing and clicking will be so small that the gap can be ignored. In preliminary research we have discovered after just one year, tweets about the Egyptian Revolution have lost approximately 11% of the resources they link to. Furthermore, content on the web is prone to change and this jeopardize the consistency of informatio conveying through time in the shared web. Recently, researchers have explored numerous social posts datasets related to specific events, topics, trends, and others. Without a way of ensuring the integrity and consistency of the shared content within these datasets we will keep on losing significant portions on a daily basis.

Temporal intention is an unexplored problem area. It exists in conventional web publishing, but is more of a problem in social media where increased volume of content, decreased textual context around each individual message, and perceived notion of disposability exacerbate the problem. In our experience, research in temporal intention proved to be difficult in large because of the lack of awareness in regards to time and how it relates to the web.

After highlighting the problem of temporal intention inconsistency we focus our research in this dissertation on the following aspects:

- Measure the general state of the resources on the web in regards to archival existence, amount persistent, deleted, or lost, measure the amount changing and have an insight on its rate and nature of change (Chapter 4).

- Analyze and model the evolution of web resources through time, from creation, sharing, editing, archiving, and (possible) disappearance (Chapter 5).

- Conduct user behavioral analysis experiments to detect what is intention in time and how to model it (Chapter 6).

- With this modeling knowledge we want to extract and analyze related features to train a classifier to model the human perception of temporal intention (Chapter 7).

- To generate the first temporal intention dataset and provide it openly for research purposes in the scientific community (Chapter 7).

- Furthermore, after properly modeling can we analyze the problem further and be able to predict the intention at t_{tweet} (Chapter 8).

- Finally, propose a framework that utilizes the developed prediction and classification models to be implemented in the form of tools to accomodate the author and the reader too to maintain the temporal consistency of the web, and enrich the archived content (Chapter 9).

We believe the lack of awareness of temporal semantics is similar to the early web phenomena of being "lost in hyperspace", but with a combination of better tools and better awareness of the idiom of browsing, users are rarely disoriented during web browsing sessions [2]. However, users do not possess the ambient awareness of time in the web in part because they do not know to ask for it. The "perpetual now" has dominated our experience for so long, most are not aware that it need not be that way. Although this has been a long-standing problem, archiving and social media tools have just now progressed where they can be combined to raise awareness about what you saw when you tweeted a link (t_{tweet}) and what your friends see when they click on it (t_{click}).

CHAPTER 2

BACKGROUND

"Dicere enim bene nemo potest, nisi qui prudenter in-
telligit." No one can speak well, unless he thoroughly
understands his subject. — Marcus Tullius Cicero

In this chapter, we briefly present the necessary terminology and definitions
that will be discussed and utilized extensively throughout the next chapters. We
introduce the anatomy of a social post and the concept of URL shortening and
aliasing which often appear in social posts linking to external shared resources. We
demonstrate the various types of web backlinks and highlight the concepts of time
and versioning on the web through public web archives and the Memento Framework.
Aided with examples and illustrations, we aim to vivify the concepts and utilities,
setting a foundation of understanding upon which we build our research.

2.1 SOCIAL POST

With the aid of social media, users can post photos, videos, personal opinions
and report incidents as they happen. With more than 1.32 billion monthly active
Facebook users as of June 2014 [3] and over 500 million tweets sent daily in 2014 [4]
social media plays a significant part in our lives. Many of the posts and tweets are
about quotidian events and the need for their preservation is debatable. However,
some are about culturally important events whose preservation is less controversial.

Social media posts differ, but they share a common ancestry and structure. To
simplify, we will be analyzing the Twitter framework and the tweet along with its
associated metadata will be the focus of our study. This study in turn could be
applied to other current or future social media forms with limited modifications.

To have a better understanding of the tweet and the parts that comprise it, we
illustrate its anatomy in Figure 10. In the rest of this proposal a social post will
refer to the textual contents of the tweet along with its publishing date, while the
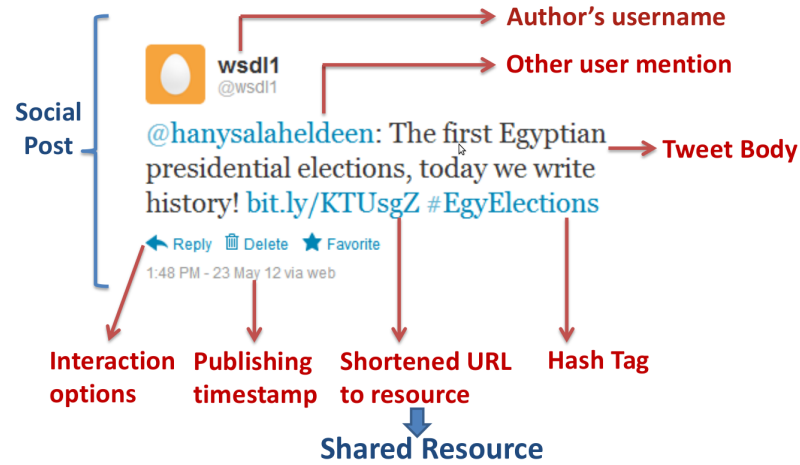
Figure 10. The anatomy of a tweet

shared resource will refer to the resource whose URI (or a shortened version of it) is mentioned in the social post.

2.2 URL SHORTENING AND ALIASING

In many cases, sharing URIs via resources has always been troublesome. Long URIs, especially ones containing parameters that can span several lines, are prone to breaking and getting cut off. Shortening a URI is a technique introduced and patented in 2000 as a method of creating a new short URI that redirects to the original long URI upon clicking the shortened one [5]. This technique has been used extensively in the last few years, especially within social networks and micro-blogging services (like Twitter) due to space constraints. In some services like Bitly, the short URLs are composed of http://bit.ly/ followed by a hash of case-sensitive, alpha-numeric string of about 1 to 7 characters. Twitter adopted automatic shortening of tweeted URIs using Bitly in 2009 and then in 2010 replaced it with its own shortening service t.co[1]. Besides shortening to avoid breaks and for space constraints, users tend to shorten URIs for various other reasons such as information hiding, tracking click logs, and ease of sharing.

Shortening is based on HTTP 30X redirects. Upon issuing an HTTP GET request to the Bitly server, for example, with the shortened URL, the server responds with a 301 Moved Permanently HTTP response with a location header pointing to

---

[1]http://radar.oreilly.com/2010/09/why-twitters-recent-announceme.html

the target URI. Then the client follows the redirection. Figure 11 illustrates a HEAD request with a follow redirects flag "-L" set to true.

```
curl -L -I http://bit.ly/losing_revolution

HTTP/1.1 301 Moved Permanently
Server: nginx
Date: Mon, 07 Jul 2014 18:19:48 GMT
Cache-Control: private; max-age=90
Location:
  http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html
Mime-Version: 1.0
Set-Cookie: _bit=53bae4c4-00328-04f10-cb1cf10a;domain=.bit.ly;expires=Sat Jan 3
18:19:48 2015;path=/; HttpOnly
Content-Type: text/html;charset=utf-8
Content-Length: 167

HTTP/1.1 200 OK
Expires: Mon, 07 Jul 2014 18:19:52 GMT
Date: Mon, 07 Jul 2014 18:19:52 GMT
Cache-Control: private, max-age=0
Last-Modified: Mon, 07 Jul 2014 18:19:07 GMT
ETag: "e3555826-b103-4daa-a3f2-d0509ebab51f"
X-Content-Type-Options: nosniff
X-XSS-Protection: 1; mode=block
Server: GSE
Alternate-Protocol: 80:quic
Content-Type: text/html;charset=UTF-8
Content-Length: 0
```

Figure 11. curl HEAD request to a bitly and following the redirects

2.3 BACKLINKS

Traditionally, a backlink refers to the link created on a page A referring to page B. Page A is considered a backlink of B. The number of backlinks could be an indication of the popularity or significance of a website or page; as well, they may be of significant personal, social, or semantic interest by indicating who is following that page. In the next sections we explore the different forms of backlinks and how we can utilize them in our investigation.

2.3.1 SEARCH ENGINE BACKLINKS

Examining the relationship between page A and page B mentioned earlier we find that it is typically straightforward to discover page B by parsing the HTML

content of page *A*. While the opposite is not that easy, discovering page *A* from page *B* is still achievable with the aid of search engines. Similarly we can utilize the search engines' APIs to discover all the pages that link to page *B*, hence we discover page *A*. It is worth mentioning that McCown and Nelson conducted a study which concluded that search engines, especially Google, under-report backlinks [6].

2.3.2 SOCIAL MEDIA BACKLINKS

Twitter enables users to associate a link with their tweeted text, technically creating a backlink to the shared resource. To illustrate, in Figure 4 the tweet about Ochlik is considered a backlink to the wordpress page in the tweet. When a user creates a web resource and publicizes it on their social network, by tweeting a link to it or posting it on their Facebook account, they create backlinks to their resource. Typically, these backlinks are not accessible via a search engine. For example, if I tweet about my personal homepage and add a link to it in my tweet, the search engines do not always extract my tweet of the homepage even though it is technically a backlink to that page. The more popular the user and the more the resource gets retweeted or shared, the more backlinks the original resource gains, increasing its rank and discoverability in search engines.

2.4 WEB VERSIONING, ARCHIVING, AND PRESERVATION

Throughout our analysis we will want to technically *freeze* the current state of the resource and store it to be utilized later. This frozen state or rather a snapshot of the resource is referred to as a *memento*. The motivation for the Memento Framework is achieving a tighter integration between the current web and remnants of the web of the past [7]. Archival versions of web resources do exist, both in special-purpose web archives such as the Internet Archive and the on-demand WebCite archive, or in version-aware servers such as Content Management Systems (CMS, e.g. Wikipedia) and Version Control Systems (e.g., Git[2], RCS[3], SVN[4], and CVS[5]). Whereas a current representation of a resource is available from its "original uniform resource identifier" (known as URI-R), prior representations - if they exist - are available from distinct resources URI-M$_i$ (i=1..n) that encapsulate the state URI-R had at times t$_i$, with t$_i$

---

[2] http://git-scm.com/
[3] http://www.gnu.org/software/rcs/
[4] https://subversion.apache.org/
[5] http://www.nongnu.org/cvs/

prior to the current time. The URI-Ms provide links back to the URI-R for which they are a memento. The resource that negotiates navigation from the current web to the past web is the TimeGate (URI-G; Figure 12). Aggregated TimeGates allow Memento clients to simultaneously access multiple archives.

In the Memento framework, the resource that provides the current representation is named the Original Resource, whereas the archival resources are named mementos. More formally, a memento for a resource URI-R (as it existed) at time t_i is a resource URI-M$_i$[URI-R@t_i] for which the representation at any moment past its creation time t_c is the same as the representation that was available from URI-R at time t_i, with $t_c \geq t_i$. Implicit in this definition is the notion that, once created, a memento always keeps the same representation. From an HTTP perspective, URI-R and URI-M$_i$ are disconnected in that HTTP provides no means to navigate towards a URI-M$_i$ via its original URI-R. Memento introduces this missing capability (Figure 12).
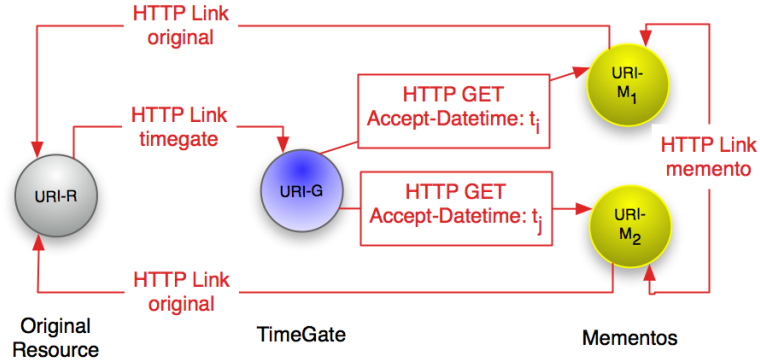
Figure 12. The Memento Framework (courtesy of Herbert Van de Sompel [1])

Inspired by Transparent Content Negotiation for HTTP specified in RFC 2295 [8] that allows HTTP clients to negotiate with HTTP servers in four dimensions (media type, language, character set, and compression), Memento introduces content negotiation in a fifth dimension, datetime. RFC 2295 introduces the notion of a transparently negotiable resource as the resource that is the target of content negotiation, and representations of that resource vary according to the aforementioned negotiable dimensions. Similarly, Memento introduces the notion of a TimeGate URI-G as a resource that supports content negotiation in the datetime dimension,

```
% curl -i mementoproxy.lanl.gov/aggr/timemap/link/http://ws-dl.
blogspot.com/

HTTP/1.1 200 OK
Date: Thu, 13 Dec 2012 03:38:36 GMT
Server: Apache
Link: <http://http://mementoproxy.lanl.gov/aggr/timemap/
link/http://ws-dl.blogspot.com>;
  rel="timemap";type="application/link-format";
  anchor="http://ws-dl.blogspot.com/"
Transfer-Encoding: chunked
Content-Type: application/link-format

Transfer-Encoding: chunked Content-Type: application/link-format

<http://http://mementoproxy.lanl.gov/aggr/timemap/link/http://
ws-dl.blogspot.com/>; rel="self";type="application/link-format",

<http://mementoproxy.lanl.gov/aggr/timegate/http://ws-dl.blogspot
.com/>;rel="timegate",<http://ws-dl.blogspot.com/>;rel="original",

<http://api.wayback.archive.org/memento/20100929000340/http://ws-
dl.blogspot.com/>; rel="first memento";datetime="Wed, 29 Sep 2010
00:03:40 GMT",

<http://api.wayback.archive.org/memento/20110202180231/http://ws-
dl.blogspot.com/>; rel="memento";datetime="Wed, 02 Feb 2011 18:02
:31 GMT",

<http://webarchive.nationalarchives.gov.uk/20120613133103/http://
ws-dl.blogspot.com/>; rel="memento";datetime="Wed, 13 Jun 2012 00:
00:00 GMT",

<http://webarchive.nationalarchives.gov.uk/20120805120725/http://
ws-dl.blogspot.com/>; rel="last memento";datetime="Sun, 05 Aug
2012 00:00:00 GMT"
```

Figure 13. A TimeMap for ws-dl.blogspot.com

and mementos URI-M$_i$[URI-R@t$_i$] as the resources that vary according to the date-time dimension. In a manner symmetrical to the way RFC 2295 introduces the `Accept-Language` request header to express the client's language preferences, and the `Content-Language` response header to express the language returned by the server, Memento introduces the `Accept-Datetime` and `Memento-Datetime` headers to express the client's preferred archival datetime for a memento, and the datetime of the memento returned by an archival server, respectively. It can be noted that, although RFC 2295 did not specify datetime content negotiation, its desirability is at least suggested by [9] as all other dimensions of genericity described in it (language, media-type, target-medium) are covered by RFC 2295.

In order to support discovery of a TimeGate URI-G for a resource URI-R, the use of a special-purpose HTTP Link header with a relationship type of *timegate* is introduced. In case of servers that have internal versioning/archiving support (such as CMS), a TimeGate URI-G for URI-R can typically be exposed by the server of the URI-R itself. In cases whereby servers rely on third parties to do their archiving (for example, by being recurrently crawled by the Internet Archive), URI-R and URI-G will reside on different servers. In addition, in order to allow discovering the Original Resource associated with a memento, another special-purpose HTTP Link header, this time with a relationship type of *original*, is introduced.

Memento also defines TimeMaps (URI-T) as a list of all URI-Ms, including the URI-R for which they are mementos and the associated datetime. TimeMaps are essentially machine-readable versions of the HTML interface. TimeMaps from aggregators sort the URI-Ms from different archives by their datetime; for example, Figure 13 is a `curl` session that returns a TimeMap (with full HTTP response headers shown for completeness) from the aggregator at Los Alamos National Laboratory (our partner in the Memento project) for our research group's blog, with two mementos in the Internet Archive's Wayback Machine and two mementos in the UK National Archives. For further technical details about the Memento framework, we refer to the original paper [7] and the IETF RFC 7089 [10].

CHAPTER 3

RELATED WORK

"If I have seen further it is by standing on ye sholders
of Giants." — Sir Isaac Newton

This work introduces the concept of *temporal intention*, and to the best of our knowledge this concept has neither been previously defined nor studied. To highlight the problem and to demonstrate our contribution, we analyze the body of work that has been done in several fields related to the problem in this section.

In this work, we are trying to detect, model, and predict user temporal intention in social media. As shown in Chapter 1, the content shared in social media is more than just kitty photos. There is a need to maintain the temporal consistency of the content shared to preserve history, and provide a better user experience during posting and reading content. In the last decade, social media and online social networks have flourished and were the focus of a multitude of studies from different angles, including recommendation, prediction, event narration and others (Section 3.1). After highlighting the significance of the shared content, we proceed to analyze the work done in the linkage between the web resources and their posts in social media (Section 3.2). The studies performed about the content itself in regards to its "aboutness", change (by alteration or disappearance) and the efforts done for change rate calculation and content replacement (Section 3.3).

We then discuss the body of work concerning the human behavior in relation to the web and social media in regards to sentiment, mood, and the various forms of intention (Section 3.4). Finally, we discuss previous possible datasets of web resources, social media posts, and links (Section 3.5) followed by an analysis of some of the crowdsourcing studies and how they were applied utilizing a service like Amazon's Mechanical Turk (Section 3.6).

3.1 SOCIAL MEDIA ANALYSIS

Due to the tremendous growth of social media [3, 4] and the continuous expansion and addition of new social network-based applications on the web [11, 12], a significant body of research has analyzed social media from many different angles.

3.1.1 UNDERSTANDING MICROBLOGGING

Microblogging, as the name suggests, is a form of user-generated communication were users can post their status, share content, or directly communicate with other users in the form of instant messages with unique identifiers. The posts are inherently short and could vary in degrees of availability from publicly posted to privately shared with specific user(s). Twitter, as discussed in the introduction chapter, is a very successful form of microblogging launched in October 2006. Other services also adopted the success of Twitter and launched their networks to provide microblogging facilities with -in several cases- a specific flavor, like Instgram (for pictures), Weibo (for the Chinese speaking community), Tumblr (for blogs and media), Identi.ca (open source), Tout (15 second videos), and others.

An early study by Java et al. attempted to analyze Twitter and acquire a better understanding of the then-new social networking phenomena [13]. They collected a dataset spanning 1.3 million posts from 76,000 users over a span of two months from April 1, 2007 to May 31, 2007. They analyzed the rate of new users joining Twitter and the growth rate of the posts published. They also examined implicit factors like the geographical distribution of users, daily trends and user communities. It is worth mentioning that this work by Java is among the pioneers in discussing the *intention* behind posting content and provided the following classifications: daily chatter, conversations, sharing information/URLs, and reporting news. They also classified the main categories of users as information sources, friends, and information seekers. Within a broad spectrum, this categorization still holds to this day.

After Java's categorization, several studies focussed on the underlying characteristics of Twitter as a form of microblogging social network. Zhao and Rosson explored why ordinary people use Twitter and its role in informal communications in a closed real network, at work [14]. In regards to content, they concluded that content shared between work colleagues tend to be more on the technological side. Also users utilize it as a form of real time people-based Rich Site Summary (RSS)

feed for the people in their respective networks.

Delving into the Twittersphere and aiming to have more in-depth understanding of the characteristics of Twitter after Java, Kwak et al. conducted an experiment where they collected 41.7 million user profiles, 1.47 billion social relations, 4,262 trending topics and 106 million tweets [11]. They analyzed the following topology and how to identify influencers in the network, a marked deviation from known human social networks as reported by Newman and Park [12]. Kwak at al. concluded that upon analyzing retweets, half of retweeting is done within an hour, and 75% in under a day while merely 10% happens a month or more after posting. After the first retweet, the tweet gets retweeted almost instantly on the second, third, and fourth hop from the original tweet, explaining the fast diffusion of the tweet.

Kwak explained that the strength in Twitter lies in the fast diffusion in comparison to Cha et al.'s report that favorite photos diffuse, or get popular, in the order of days on Flickr [15]. Also Yang and Counts analyzed the speed, scale and range of the posted content on Twitter to have a better understanding its information diffusion patterns [16].

With content being shared around the clock, researchers have addressed the concept of trending topics in social media. Cataldi et al. developed an approach to detect in real-time emerging topics on Twitter by extracting the tweets' contents and modeling the extracted terms' life cycle by the use of an aging theory to extract emerging terms which map to topics in user-specified time frames [17]. Chen et al. analyzed topic detection as well [18], while Weng and Lee analyzed event detection in Twitter [19]. Mathioudakis and Koudas on the other hand analyzed trend detection in Twitter stream [20], as did Benhardus and Kalita [21]. As for trending news, Phuvipadawat and Murata analyzed the Twitter stream and focused on tracking these news in real-time [22]. Recently, Xie et al. developed TopicSketch, which is a tool for bursty topic detection in real-time from the Twitter stream [23].

Beyond textual tweets, image tweets have been reportedly increasing in importance and popularity in social networks. Yu et al. reported that almost 56% of the microblog posts on Weibo were image tweets in 2011 [24]. Also image tweets have a higher retweeting rate and longer survivability [25]. Chen et al. explored image tweets to have a better understanding of the classification of the visually relevant and non-relevant images in textual tweets [26].

3.1.2 HISTORY NARRATION

In chapter 1 we illustrated that social media is not used on a daily basis for merely status updates, what the user had eaten today, or their pet pictures. In a multitude of cases it is utilized in conveying worthy information, event narration, or broadcasting time-sensitive information. This posted and shared content in relation to a current event could be utilized by future scholars as a collective narration of the thoughts, vibe, interactions, and perception of the people in relation to that event.

In October 2010, Malcolm Gladwell, a writer for the *The New Yorker*, wrote an article arguing that the role played by the social media like Facebook and Twitter in relation to protests and revolutions has been highly exaggerated [27]. He argues that the poor revolutionary power of the social networks is because they encourage lazy activism by merely clicking a button instead of getting out and doing something. In his article he pointed out how events unfolded in the early 1960s, leading to a civil rights movement that spanned a decade. Within the same week of publishing the article at *The New Yorker*, another writer named Leo Mirani published an article in *The Guardian* opposing Gladwell's opinion [28]. He illustrated the power of social media in aiding protests by an example from the Kashmir protests, in the same summer of 2010, which gained a lot of momentum and worldwide coverage in the press, and how it is correlated to the increase in social media users in the region. He finally opposed the definition of "activism" from Gladwell's prospective that contemporary activism might surpass just going out to the protest to actively sharing, posting, and changing people's minds on a large scale.

Starbird and Palen also called Gladwell's claim into question by analyzing the retweeting mechanism on Twitter to reveal the aspects of "work" that the crowds conducted to diffuse information in relation to the 2011 Egyptian Revolution [29]. They analyzed the tweet content during a mass emergency and were able to identify the locals who authored the original content that was retweeted and also measure the authors' interaction and reaction to that content. This study followed an earlier experiment where they analyzed Twitter communications during the flooding of the Red River Valley in the US and Canada in 2009 [30]. To enhance the situational awareness during a crisis and to aid building working software systems and frameworks to be used by the first responders and the public, Starbird utilized the data collected from the Red River Valley incident along with data collected from the Oklahoma Grassfires in 2009 to identify features of the information generated by

the masses to be utilized in building the goal frameworks [31].

Other studies have analyzed social media content in relation to world events or crises and the use of this shared content in dealing with those disasters. Qu et al. investigated the 2010 Yushu earthquake in China and how microblogging (in this case via Weibo) was utilized to broadcast immediate needs and solicit donations [32]. Neubig et al. analyzed the Twitter content related to the victims of the 2011 East Japan earthquake to mine for information safety and extract it robustly then deliver it to the affected people in the area [33].

Opposing Gladwell, Starbird and Palen also examined the motivations, resources, activities, and products of digital volunteers on Twitter (or voluntweeters) to analyze how they self organized during the 2010 Haiti earthquake [34]. Finally, to address Gladwell's claim that "high-risk activism fails in social media", we mention Burns and Eltham's work which argued that the online services may be used in some cases by the government to crush opposition protests and identify protesters and thus jeopardize their lives [35].

As for the shared content and posts themselves, it is highly beneficial to be able to identify the sources of the social activity online during disruption events. By being able to narrate the events play by play from the ground during a mass disruption, on-the-ground tweeters have higher legitimacy. Starbird et al. conducted an experiment to build two classifying models based on a dataset collected during the 2011 New York Occupy Wall Street protests [36].

From a different angle, Lehmann et al. attempted to identify news curators among the mass of daily tweeters [37]. A news curator is an individual that exerts a substantial amount of effort to monitor a large variety of sources on a topic or around a story and extract the contents related to the desired topic and disseminate it to the public. This identification and classification of a specific group of users increases the probability of obtaining legitimate and credible news. Lehmann also devised a method of defining transient news crowds to help journalists and news editors to rapidly detecting followup stories to their published articles, thus increasing the awareness of the evolution and propagation of published content [38]. Finally Mark et al. analyzed the long term effects of disaster events or wars on the longer lasting content of social media, namely blogs [39], while Gill et al. analyzed the motivation and topicality of the published blogs [40].

3.1.3 SOCIAL ANNOTATIONS

Web annotation is a form of online annotation associated with a web resource, typically a web page. Users can add, alter or remove information from a Web resource in separation from the resource itself. This user generated content is typically uncontrolled and volunteered, thus it is called social annotation. Several online services are based mainly on the concept of social annotation by the masses, as it takes several forms like tags, likes, comments, bookmarks, pins and others.

With the emerging phenomenon of social annotations, research has been done to investigate the value of such tags for search in the web. Bao et al. observed that tags from del.icio.us are usually good summaries of the corresponding web pages and the count of the tags indicates the popularity of the pages [41]. Social annotations from del.icio.us utilized in enhancing web search were also exploited by Yanbe and Jatowt et al. [42]. They propose to combine the current link-based ranking methods with characteristics derived from social annotations and introduced SBRank which captures the popularity of a page. SBRank is computed by counting the number of times a page has been bookmarked (voted for by users) and can therefore be seen as a simplistic version of Social Page Rank (SPR) as presented in the works of Bao et al. mentioned above [41]. The authors implemented a prototype search portal, which enables searching by common query terms as well as by tags. The user can also give certain weight to the source, e.g. have tags twice as important for the query as the common terms. The ranking of the results is determined by combining link-based methods and the output of SBRank.

Heymann and Koutrika et al. investigated the relationship between tags and the web pages they refer to (taken again from del.icio.us) as well as the tags and their URLs compared to the query terms and URLs from the AOL search logs [43]. Roughly 9% of the top 100 results for search queries (from the AOL logs) are annotated in del.icio.us, and this coverage doubles to 19% when considering only the top 10 results. That means despite the relatively small coverage of web pages, del.icio.us URLs are disproportionately common in search results. They also found that tags significantly overlap with popular search terms which indicates that tags can indeed help locating relevant pages. Interestingly, despite the overlap, tags and search terms were not correlated: 50% of the tags annotating an URL either occur in the text of the page itself and 16% of the tags even occur in the page title. Astonishingly, 80% of the tags occur in either the page they refer to or in

one of the in- or outlinked pages. They found the vast majority of the tags to be relevant to the pages they refer to and also that tags are often highly correlated with particular domains and vice versa. Furthermore, Heymann studied other aspects of social media annotation or tagging in regards to prediction [44], human knowledge [45], expert analysis in the process of tagging [46], and the use and abuse of the tagging data in a collaborative environment [47].

Alongside del.icio.us, Bischoff et al. investigated last.fm (a music portal) and Flickr (a photo portal) in their tag category analysis [48]. Due to the variety of the sources, they classify the tags into eight main categories. Different categories are important for different domains, e.g., the category "topic" was dominant for tags from del.icio.us and Flickr since it describes the domain and anything that can be seen on a picture, but the category "type" was prominent for last.fm tags since it describes the file format as well as the music genre. Therefore, the predicted usefulness of tags for web search (assessed by a user study) depends on the category of the tags. This observation is intuitively confirmed since tags that belong to the "location" category are more useful to discover an image on Flickr than music from last.fm or a bookmark from del.icio.us. Of the total number of tags obtained from del.icio.us, 44.85% were occurring in the text of the annotated page, and this shows that more than 50% of the tags provide new information about the URL they describe. This extra information could be utilized for web search. Klein and Nelson introduced the notion of ghost tags, which they used to describe terms used as tags that do not occur in the current but did occur in a previous version of the web page [49].

Social annotations in resource discovery are useful on a personal level as they are essentially markings indicating that a person in the social network of the user has liked or shared a specific document from the list of results of a query that the user issued. The user can benefit from such social experiences in various ways, including the discovery of socially vetted recommendations, personalized search results, and emotionally connecting with an otherwise static and impersonal search engine. Pantel et al. devised a taxonomy of aspects that influence the perceived utility of social annotations in a Web search scenario, drawn from the query, social connection, and content relevance [50].

3.2 LINK ANALYSIS

A large percentage of the social content posted in social media contains a link to an external resource by including a URL in the post. In the SNAP dataset of tweets, which we will describe in detail in section 3.5, approximately 38% of the randomly collected tweets in it had an embedded URL linking to an external resource in 2009, and this percentage is increasing. This external resource could be a text-based web page or a media file like an image or a video. The purpose of this inclusion is to enhance the posted content, provide supporting evidence, extending the story, or other.

Researchers have investigated the automatic generation and inclusion of links to enhance the content and in several cases to maintain user engagement. To illustrate, an example is to automatically generate links to encyclopedic content to enhance the knowledge in the document [51, 52]. Automated entity linking and discovery is also analyzed from multiple prospects among which: human intelligence augmentation [53, 54]. Link generation is also utilized in disambiguation tasks [55, 56, 57]. Automatically enriching articles with news worthy links has also been investigated by Ceylan et al. [58]. They propose a new automated system that detects newsworthy events without relying on resources like Wikipedia to identify those events. This is because in several cases, Wikipedia will not contain information about very recent contemporaneous *newsworthy* events. Their system was designed to function independently from the analyzed domain; and this system was evaluated using Amazon's Mechanical Turk.

Several studies have focused on the relationship between the posted URI and the content of the intended resource. Klein and Nelson proposed building a framework for describing the mapping between the URIs and content [59]. They defined four different scenarios of the relationship between a URI and a resource's content as follows: the same URI maps to the same or very similar content at a later time, the same URI maps to a different content at a later time, a different URI maps to the same or very similar content at the same or at a later time, and finally the content can not be found at any URI.

Finally, it is worth mentioning that authors sometimes tend to shorten the links to their articles using one of the logging shortening services like Bit.ly to closely analyze the resulting click-logs to gauge their audience's interaction and dissemination of the articles [60].

3.2.1 URL SHORTENING

Shortened URLs are normally created to replace long ones (as shown in section 2.2 and the HTTP response in Figure 11) to ease dissemination and to solve half a dozen other problems [5]. Several research papers addressed the aspects of the URL shortner implementation or solving its problems which might occur, like linkrot, privacy issues, blockage, and others. Unfortunately, and to the best of our knowledge at the time of writing this dissertation, one one study addressed the concept of short URLs in the field of social media. Antoniades et al. studied the use of URL shorteners, especially with respect to their use in social media [61]. In that study, they argue that short URLs are not ephemeral, with roughly 50% active for more than three months, and they emphasize the fact that short URLs reflect an "alternative" web.

3.2.2 BROKEN LINKS AND LINK ROT

URLs are always prone to change due to the dynamic nature of the web making the durability of the published URLs a necessity in multiple cases. To ensure this, Tim Berners-Lee published a set of guidelines for creating durable URLs [62]. Durability or "coolness" means that URIs should not change based on date of access, representation, or how the webpage is structured to various users. Unfortunately there is a lack in link integrity on the web, as demonstrated by Ashman and Davis in their respective works [63, 64, 65, 66].

Koehler conducted a very interesting four-year longitudinal study that concluded that a random test collection of URLs eventually reached a steady state, after approximately 67% of the URLs were lost over a 4-year period, and thus estimated the half-life of a random web page is approximately two years [67].

In 2000, Lawrence et al. concluded that between 23 and 53% of all URLs occurring in computer science related papers authored between 1994 and 1999 were invalid [68]. By a manual multi-level search on the Internet, they were able to reduce the number of inaccessible URLs to 3%.

Spinellis conducted a similar study investigating the accessibility of URLs occurring in papers published in Communications of the ACM and IEEE Computer Society [69]. They found that 28% of all URLs were unavailable after five years and 41% after seven years. They also found that in 60% of the cases where URLs where

not accessible, a 404 error was returned. They estimated the half-life of an URL in such a paper to be four years from the publication date.

Focusing on articles published in the D-Lib Magazine, McCown et al. showed that the average half-life of these articles is 10 years [70]. While in the field of digital libraries, Nelson and Allen studied object availability in digital libraries and found that 3% of the URLs were unavailable after one year [71]. Loss of references and URIs appearing in the academic literature have been studied numerous times, with exact loss rates varying depending on the corpus [72, 73]. While analyzing the availability of web resources referenced from papers in two scholarly repositories, Sanderson et al. discovered a startling 45% of the URLs referenced from arXiv still exist, but are not preserved for future generations, and 28% of resources referenced by UNT papers have already been lost [74]. In similar scholarly context, Klein et al. analyzed a dataset of 3.5 million articles and discovered that an average of one in five scholarly articles suffers from reference rot [75].

Internet references were examined by Dellavalle et al. in articles published in journals with a high impact factor (IF) given by the Institute for Scientific Information (ISI) [76]. They found that Internet references occur frequently (in 30% of all articles) and are often inaccessible within a month after publication in the highest impact (top 1%) scientific and medical journals. They discovered that the percentage of inactive references (references that return an error message) increased over time from 3.8% after 3 months to 10% after 15 months up to 13% after 27 months. The majority of inactive references they found were in the .com domain (46%) and the fewest in the .org domain (5%). By manually browsing the IA they were able to recover information for about 50% of all inactive references.

A similar study was conducted by Markwell and Brooks, observing links from a Biochemistry course intended for distance learning for high school teachers [77]. They also found that the number of accessible links steadily decreased, and after one year 16.5% of their links were non-viable. They observed that the .gov domain was the most stable one, and links referring to the .edu domain were more transient. Of these links, 17.5% had disappeared within a year.

The problem of disappearing or changing resources has also been well-studied throughout the last decade. The aspect of web decay has been analyzed by Bar-Yossef et al. [78] and they proposed a measure of decay and algorithms to compute it efficiently. They also realized that not only single web pages but collections and

even entire neighborhoods of the web show significant decay.

3.2.3 ROBUST LINKS

Given the broken nature of the links, several attempts were done to recover from this change, loss or failure. System administrators on the requested server perform redirects via response code 30X when they are aware of the change in location of the resource on their server.

Another approach is to adopt more permanent identifiers that are more persistent and act as an intermediate. A Digital Object Identifier (DOI) is a permanent identifier of an digital object that can be resolved to an instance of the required data [79]. The DOI is resolved through the Handle system [80]. PURL (Persistent Uniform Resource Locator) does not refer to the location of the resource itself but to a (supposedly) more persistent, intermediate location through HTTP redirects [81]. PURL is used to redirect to the location of the requested web resource where it redirects HTTP clients using HTTP status codes. PURLs are used to curate the URL resolution process, thus solving the problem of transitory URIs in location-based URI schemes like HTTP [82].

Nakamizo et al. developed tool that discovers the new URL of a web page in case it has been moved. The link authorities are reliable web links that are updated as soon as a pages moves [83]. Furthermore they enhanced the Pagechaser tool with heuristics based on assumptions about the location of the page that has been moved by using HTTP redirect information if available and performed a keyword search with web search engines to locate the new page [84].

Errorzilla is introduced as a browser extension to the Mozilla browser project that implements a useful error page when a website cannot be reached [85]. It adds *Try Again, Google Cache, Coralize, Wayback, Ping, Trace,* and *Whois* buttons, along with the Firefox logo to the error page when a website is not found or a web server is down.

3.3 SHARED CONTENT ANALYSIS

The third and last component of the process of sharing content in social media is the content itself. After analyzing the social aspect and the linkage, we analyze the content itself in the external resource. As most of the resources that are available in the public web, it is vulnerable and prone to change, or loss. Firstly, to

understand why that external resource was incorporated in the social post we discuss the concept of "aboutness", which defines what this resource is about or what is its subject/topic. Aboutness is a term used in library and information science (LIS), linguistics, philosophy of language, and philosophy of mind. In LIS, it is often considered synonymous with subject (documents). In philosophy it has been often considered synonymous with intentionality[1]. Secondly, when this aboutness change from the original state when it was first mentioned in the social post, this indicates a content change. Thus, we will investigate next the content change, the rate it changes, and its possible decay and disappearance. Finally, we investigate the possibility and feasibility of finding possible replacements for the changed or disappeared content.

3.3.1 ESTIMATING ABOUTNESS

Aboutness in this context is a form of topic detection with a broad spectrum of non-predefined topics. In this section we will investigate TF-IDF and Lexical Signatures as two methods of extracting a specific set of terms from the content of a document capturing its aboutness.

TF-IDF

In defining the aboutness of a page, for the first step, stop words need to be identified and eliminated in the document. Wilbur and Sirotkin introduced a method to automatically identify stop words in a given corpus [86]. Their claim is that stop words have the same probability to occur in both documents not relevant to a given query and documents relevant to the query. Stop words are often eliminated from the documents via a stop word list and the remaining terms are usually shortened to their stems (both language dependent) in order to avoid quasi duplicates due to trivial word variations. Probably the most famous and commonly applied stemming algorithm is the porter stemmer, first introduced by Porter [87].

On the one hand, term frequency (TF) answers the question, "how often does a specific word appear in a certain document?" It justified by the probability that a term that occurs very frequently in a document is likely to be more relevant for that document than a term that occurs less frequently. Term frequency is rather trivial to compute since it only depends on the number of terms that occur in a

---

[1] http://en.wikipedia.org/wiki/Aboutness

document. However, since documents vary in length and therefore in number of terms, TF values need to be normalized as demonstrated by Singhal et al. [88].

On the other hand, inverse document frequency (IDF) answers the question "in how many documents does a specific word appear?" IDF is a measure of term specificity as discussed by Jones in 1972 in the context of improving automatic indexing for retrieval systems [89]. Furthermore, Robertson provided theoretical arguments for the good performance of IDF [90] and a method to measure the global importance of terms [91].

IDF depends on knowledge of the entire corpus. In particular the IDF computation requires knowledge about the total number of documents in the corpus and the number of documents the term occurs in. Salton et al. presented a good overview of TF-IDF as a term weighting approach in text retrieval (and automatic indexing) [92, 93].

The TF-IDF scheme is used to represent the content of web pages without particularly focusing on the lexical signatures. Sugiyama et al. claim that for documents in a hyperlinked structure like the Internet the content of neighboring pages need to be exploited too, in order to obtain more accurate descriptions of a page [94]. Their research is based on the idea that the content of a centroid web page is often related to the content of its neighboring pages. Topical locality has also been analyzed by Davidson [95].

Dean and Henzinger defined neighboring pages as pages that refer to the centroid page (inlinks for the centroid) and pages the centroid links to (outlinks) [96]. They show that by refining the original TF-IDF with input from the neighborhood the performance of the lexical signature (which will be explained in the next section) in terms of precision and recall while querying search engines for related pages can be improved.

Lexical Signatures

In our research we will utilize what we call "tweet signatures" in reconstructing missing web content as illustrated in the following chapter. This tweet signature is based mainly on a data mining technique named *lexical signature.*

A lexical signature is a small set of terms derived from a document that capture the "aboutness" of that document. It can be thought of as an extremely lightweight metadata description of a document, as it ideally represents the most significant

terms of its textual content. Phelps and Wilensky first introduced the term lexical signature (LS) and proposed their use to discover web pages that had been moved and confirmed that 5-term LSs are suitable for discovering a page when used as search engine queries [97]. In absent resources within a digital collection, and with also no valid metadata associated to the missing resource to be found, Meneses et al. explored the viability of using the lexical signatures of valid documents within a collection to find suitable replacements for absent resources [98].

Lexical Signatures are usually generated following the TF-IDF weighting scheme which gives each term a significance weight within the collection of documents. There are limitations on Phelps and Wilensky lexical signatures though. Their scenario required the browser's source code to be modified to exploit LSs and they required LSs to be computed a priori. Park et al. studied the performance of nine different LS generation algorithms (retaining the 5-term precedent) and proved that slight modifications in the generation process can improve the retrieval performance of relevant web pages [99].

Wan and Yang devised another method for lexical signature generation based on the "WordRank" [100]. Their method takes the semantic relatedness between terms in a LS into account and chooses the most representative and salient terms for a LS. The authors also examined 5-term LSs only and found that DF-based LSs are good for uniquely identifying web pages and hybrid lexical signatures (variations of TF-IDF) perform well for retrieving the desired web pages. They claim, however, that WordRank-based LSs perform best for discovering highly relevant web pages in case the desired page can not be located.

Staddon et al. devised a lexical signature-based method for web-based inference control [101]. Following the TF-IDF method, they extract salient keywords (which can be considered a LS) from private data that is intended for publication on the Internet and issue search queries for related documents. From these results they extract keywords not present in the original set of keywords, which enables them to predict the likelihood of inferences. These inferences can be used to flag anonymous documents whose author may be re-identified or documents that are at risk to be (unintentionally) linked to sensitive topics.

Another form that could be thought of as a lexical signature is Henzinger et al.'s work in generating related web pages to TV news broadcasts using a two-term summary [102]. This summary is extracted from closed captions and various

algorithms are used to compute the scores determining the most relevant terms. The terms are used to query a news search engine where the results must contain all of the query terms. The authors found that one-term queries return results that are too vague and three-term queries return zero results too often, thus they focus on creating two-term queries.

Klein et al. utilized lexical signatures extensively in identifying content about-ness, and rediscovering content on the web [103, 104, 105].

3.3.2 CONTENT CHANGE

The web is ever-changing and what one might share or post today might change or disappear tomorrow. Losing web resources and finding them again has been the scope of several studies. In this section we explore the methods for detecting the change or decay in content, quantifying it, and finally the attempts to replace it.

Detecting Change

As the content on the web changes, it is crucial to detect and quantify this change to have a better understanding of the evolution course of the page and the type of this change. Changes differ in type and significance according to the type and structure of the resource. In this section we discover some of the previous works in detecting and quantifying change in published content.

To compare two web pages or two versions of the same page, we can utilize hash comparisons. If we consider a web page as the input to the hash function, we can compare its output to the output of the hash function of the other page. Furthermore, since a lexical signature captures the content of the page, and ignoring the fact that the hash value is the transformation of the entire content and the lexical signature consists of a limited number of significant terms only, we can compare the output of the hash function when we provide the lexical signatures of the two pages as input. Changes in the input set of the web page are reflected in the hash value, as long as they are significant and also in the lexical signature.

In common hash functions such as MD5 Message-Digest Algorithm [106] and US Secure Hash Algorithm 1 (SHA1) [107], the output changes dramatically, even with the smallest changes in the input set. Charikar introduced the Simhash function, which is different from other hashing functions [108]. For any given input set, the Simhash function changes relative to the modification of the input. That means if

the input only changes slightly, the change in the hash functions is minor, and if the input set changes significantly, the change in the hash function is major. Simhash can be applied to find similar web pages in order to improve the quality of a web crawler [109].

For digital libraries, Nelson and Allen analyzed the persistence and availability of objects in a digital library [71]. In web archiving, avoiding unnecessary downloads of unchanged pages can significantly reduce the load on both the archiving system and the server being archived. Thus it was crucial to detect the content change and illustrate a scheme for reliably predicting whether content has changed without having to download the content. Clausen utilized Etags and last-modified date fields to achieve this prediction [110]. He sampled the front pages of all Danish second-level domains and for each page, he recorded the date, the Etag, the size, and an MD5 sum of the body of the page. He illustrated that over 80% of the downloads done in this experiment could have avoided if an accurate predictor of content changes had been available. He also concluded that frequently changing pages tend not to have Etags and the Etag header is missing in 40% of all downloads, while the Last-Modified header would give errors in 0.30% of all changed pages but would avoid 63.7% of unnecessary downloads.

The changing aboutness of live web pages has been studied in the Walden's Path project [111, 112]. Walden's Paths' Path Manager is a tool that allows users to construct trails or paths using web pages, which are usually authored by others. The path can be seen as a meta-document that organizes and adds contextual information to those pages. Simply comparing the candidate page with a cached copy may not be sufficient for them because some changes are actually desirable and should not be automatically dismissed. It is possible that pages change on a constant rate (such as weather or news sites) and therefore a simple comparison is not sufficient. Their focus, however, is on discovering significant changes to pages, and their evaluation of change is based on document signatures of paragraphs, headings, links and keywords. They also keep a history of these values so that a user can actually determine long-term as well as short-term changes.

In regards to the frequency of content change, Adar et al. concluded that web resources that change more frequently are shown to contain more important content [113]. Finally, identifying the rate of change and computing it for various web resources is a well-studied phenomena. To examine the estimated frequency of

change and the quantification of this change or loss, Cho studied the change rate of web pages to determine the best policies for web crawlers [114, 115], as well as studying how to handle late arrivers in a collection [116]. Other studies have been done about detecting change as well [117, 118, 113, 119]. Crawl policies for enhancing archival coverage have been studied too [120, 121, 122, 123].

Soft 404s

A soft 404 page is a page returning actual page not found errors, but instead of returning the HTTP response code 404, it returns code 200 (meaning OK). Figure 14 highlights an example of a soft404 page.

An approach to learn whether or not a web server produces soft 404s is achieved by Bar-Yossef et al. by sending two requests to a suspicious server [78]. The first request is asking for the page of interest and the second for a page that with very high probability that it does not exist. It then compares the server behavior for the two returns, such as number of redirects. The content of the returned pages are also compared so that, in case the two behaviors and the content of the returned pages are very similar, the algorithm gives a clear indication of having detected a soft 404.

Soft 404s usually return pages that the user did not expect, thus the difference between the expected page (the one the user has experienced before) and the actually returned page is significant. Sometimes a domain gets reregistered. Its content has continued to evolve beyond the intentions of the original maintainers. A malicious example to this is domain hijacking, where a party steals the domain in order to distribute their content. The domain sex.com is probably the most famous example of domain hijacking as illustrated by Ramasubramanian and Sirer [124] and the ICANN Security and Stability Advisory Committee SSAC [125].

3.3.3 CONTENT REPLACEMENT

We established that content within a page, or the whole page, or even entire websites disappear on a regular basis for a multitude of reasons. In this section we explore the research done on missing content replacement and recovery. McCown et al. argued that various reasons were found for why entire websites go missing and how they potentially can be recovered [126]. Also in his doctoral dissertation, McCown presented extensive research on the usability of the web infrastructure for reconstructing missing websites from the web infrastructure [127].
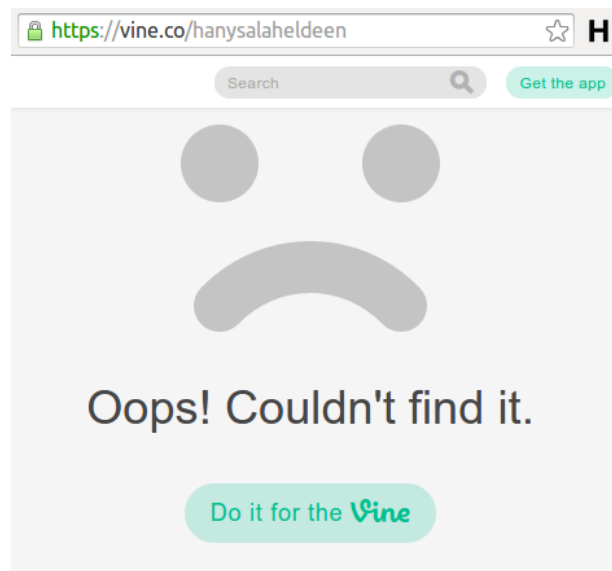
(a) The vine page for a none-existing page

```
$ curl -I -L https://vine.co/hanysalaheldeen
HTTP/1.1 200 OK
Cache-Control: max-age=3600
Content-Type: text/html; charset=utf-8
Date: Fri, 17 Apr 2015 10:42:11 GMT
Strict-Transport-Security: max-age=631138519
X-Content-Type-Options: nosniff
X-Frame-Options: SAMEORIGIN
X-XSS-Protection: 1; mode=block
Connection: keep-alive
```

(b) HTTP response headers for the same non-existing page

```
$ curl -I -L https://vine.co/blablabla
HTTP/1.1 200 OK
Cache-Control: max-age=3600
Content-Type: text/html; charset=utf-8
Date: Fri, 17 Apr 2015 10:42:22 GMT
Strict-Transport-Security: max-age=631138519
X-Content-Type-Options: nosniff
X-Frame-Options: SAMEORIGIN
X-XSS-Protection: 1; mode=block
Connection: keep-alive
```

(c) HTTP response headers for the a manufactured "blablabla" page that we know it does not exist on the server

Figure 14. Last modified date example

Consequently, several memebers of our Web Science and Digital Libraries research group (WS-DL) analyzed the loss and rediscovery of websites to pin point the reasons behind this behavior [128, 129, 130]. Furthermore, they investigated a variety of techniques, including using page titles [131], tags [49], and lexical signatures [104, 132, 133], all of which could be used as queries to search engines to find replacement copies of the missing web page.

In our Web Science and Digital Libraries group, we presented three other systems to reconstruct missing websites by finding missing web pages or their alternatives as follows:

Opal: Harrison introduced Opal as a system that feeds LSs into the Web Infrastructure to find missing web pages [134, 135]. The main difference here is that Opal is a server side system and therefore requires system administrators to install and maintain the software.

Warrick: McCown introduced Warrick as a system that implements "Lazy Preservation" [136]. Warrick crawls web repositories such as search engine caches and the index of the IA to reconstruct websites. His system is targeted to individuals and small scale communities that are not involved in large scale preservation projects and suffer the loss of websites.

Synchronicity: Klein introduced Synchronicity as a system that locates the missing page or sufficient replacement pages in real time [59]. It uses information retrieval techniques (like LSs) to (re-)discover the pages and recovers a single resource (a web page) at a time. Synchronicity is geared towards end users, browsing the web and experiencing HTTP 404 errors.

3.4 HUMAN BEHAVIOR ANALYSIS

Intention, mood, and sentiment have been analyzed in different contexts, but none with respect to time. This research builds on a large body of work involving the different aspects of human behavior, specifically the temporal intention. To highlight the differences we examine the previous works in related fields of sentiment, mood, and intention.

User behavior in general has been studied numerous times. Benevenuto et al. studied the user workloads in online social networks [137]. They conducted a 12-day data collection analysis summarizing HTTP sessions of 37,024 users in Brazil who accessed four main social networks: Hi5, Orkut, Myspace, and LinkedIn. They also

presented a clickstream model to characterize user behavior in social networking websites. Also Dupret and Lalmas conducted several studies to gauge and measure user engagement during their experience on a website by analyzing time between visits (or absence time) [138], keeping users on the website longer by providing them enhanced clickthrough experiences where they navigate links - on the same website - to newsworthy events [58], measuring the inter-site engagement for users navigating through the partner websites of an entire content provider network [139, 140] and the effect of links within this network [141], and finally, modeling this user engagement [142].

3.4.1 SENTIMENT

Generally, sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall contextual polarity of a document[2]. The attitude may be their judgment or evaluation, affective state (that is to say, the emotional state of the author when writing), or the intended emotional communication (that is to say, the emotional effect the author wishes to have on the reader).

In many cases, sentiment and intent go hand-in-hand in analyzing social networks interactions and posts in the blogosphere. Mishne and Glance analyzed the sentiment in weblog posts to predict movie sales [143]. Durant et al. succeeded in predicting political sentiment by analyzing web logs correctly with an average of 89.77% using a Naïve Bayes classifier coupled with feature selection [144, 145].

Kucuktunc et al. conducted a large scale analysis on the effect of sentiment on the answers posted on Yahoo! answers [146]. They also showed that the sentiment in the answer is correlated to its selection of being the best answer. In financial and business topics, best answers often have more neutral sentiment than other answers. Also they were able to predict the attitude that is provoked in the answers, thus understanding the factors affecting the collective mood and linking it to sentiment. This could be utilized further in advertising, search, and recommendation tasks.

There has been a significant progress recently in sentiment analysis and gauges for public and individual mood, especially using Twitter feed and blog content. Measuring emotions and sentiments have ranged from measuring happiness and worry [147, 148], and sentiment analysis was also employed in several applications from

---

[2]http://en.wikipedia.org/wiki/Sentiment_analysis

predicting elections [149, 150], to news recommendations [151], to sales prediction [152]. Furthermore, Twitter was utilized as a corpus for opinion and sentiment mining [153], sentiment identification in events, [154] and in general sentiment prediction [155].

3.4.2 MOOD

In psychology, a mood is a temporary emotional state[3]. Moods differ from emotions or sentiment in that they are less specific, less intense, and less likely to be triggered by a particular stimulus or event. Moods generally have either a positive or negative valence. In other words, people typically speak of being in a good mood or a bad mood. Mood also differs from temperament or personality traits which are even longer lasting.

Social media, and Twitter specifically, has been analyzed in regards to the collective mood of users and how this mood transitions over time as observed in the public timelines. Mogadala and Varma investigated the mood transition phenomena while analyzing user collective behavior and was able to successfully predict this mood transition [156]. Bermingham and Smeaton analyzed and monitored collective political mood and sentiment on Twitter and argued its viability in predicting the election results [157]. Bollen et al. analyzed the textual content of the daily Twitter public feed and applied OpinionFinder (a publicly available software package for sentiment analysis) which measures positive and negative mood; and Google's Profile of Mood States (GPOMS) which predicts one of six mood dimensions (Calm, Alert, Sure, Vital, Kind, and Happy) to successfully predict the stock market DJIA with 86.7% accuracy in the daily up and down changes in closing values [158, 159, 160].

Another form of mood analysis is in the field of music classification, in which raters classify songs according to the mood that best represents the song they are hearing [161, 162]. Furthermore, several studies focused on measuring and defining the global mood levels in blog posts, among which, the work conducted by Mishne and Rijke [163].

3.4.3 INTENTION

Although user intention has been widely studied, it has only been applied to the area of web search, e-commerce, web spam detection, and political and economical

---

[3]http://en.wikipedia.org/wiki/Mood_(psychology)

sentiment analysis. To the best of our knowledge, it has not been applied to the temporal intention of users and the bridge between the current and past web. At the time of writing this document, there is no published research describing temporal intention in the context of web navigation and social media dissemination.

User intent has been studied, analyzed, and predicted in several works in the past decade. These works span multiple fields ranging from psychology, sociology, computer engineering, to computer science. Focusing on the latter field, user intention has been addressed from different angles.

Researchers have studied and analyzed the user intent behind queries in web search [164, 165, 166, 167, 168, 169, 170, 170]. Other studies focused on understanding users' click models for query intent [171, 172, 173]. Santos et al. utilized intent analysis in search result diversification [174].

Na Dai et al. proposed classifying the intent expressed by web content creators and classified it as navigational or informational [175]. The same authors published a follow-up study to bridge the gap between the link intent and the query intent, and how this gap filling will enhance web search quality [176].

User intention has also been studied extensively in the commercial field. Guo and Agichtein analyzed the relationship between search intent, result quality and searcher behavior in online purchases and how optimizing these interactions can enable more effective detection of searcher goals [177]. Furthermore, commercial intent analysis was used in web spam detection and resulted in improving the spam classification by 3% [178]. Intent analysis is also utilized in spam and phishing attacks detection [179, 178].

As for the temporal aspect of intention analysis, Zhou et al. analyzed the effect of temporal intent variability in diversifying search results [180]. To cope with the uncertainty involved with ambiguous or underspecified queries, search engines often diversify results to return documents that cover multiple interpretations. The temporal subtopic popularity change is common for many topics, and they concluded that temporal subtopic popularity variability is modest or high for over 35% of ambiguous topics, and has considerably significant impact on diversity evaluation.

Furthermore, intention analysis and detection in web science have several variations and can be found in different contexts. It was analyzed as an independent concept [181, 182, 183], in cluster analysis [184], gaming [185], energy management [186], in data mining [187], in microblogging [13], as well as in psychology [188, 189].

3.5 DATA COLLECTIONS

For the purpose of this study we will collect, analyze, and utilize various datasets of web pages, archived content, click logs, and social posts. Several research groups and affiliations have released various types of datasets for research purposes during the last few years. The Text REtrieval Conference (TREC), which is co-sponsored by the National Institute of Standards and Technology (NIST) and the U.S. Department of Defense, have released through their annual TREC Web Track a dataset, which consists of 870-million of web pages crawled over the last years to be utilized in specific competitive tasks annually [190]. The TREC collection consists of a finite (and well known) number of pages and the textual content of all pages is available, researchers can compute IDF values for terms. This collection is assumed to be a representative sample of the entire web as discussed by Soboroff [191]. Due to the dynamics of the Internet however, Chiang et al. claimed that TREC collections are dated and results from TREC-based data can not be considered accurate when working with web page content [192].

Stanford's Network Analysis Project (SNAP) have released several datasets collected from social networks, communities, wikipedia networks and meta data, online reviews, Twitter posts, Memetracks, Facebook networks, and several others, to be utilized for research purposes in 2009 [193]. In the next chapter we will highlight our usage of their Twitter network dataset of 476 million tweets.

3.6 CROWD SOURCING

As we illustrated in Section 3.4, the task of modeling human intention is highly subjective and requires human intelligence in order to correctly assign the desired classification to the ground truth data set. To perform this on a large scale, we utilized Amazon's Mechanical Turk to perform this assignment. Within the last couple of years, researchers from various fields utilized Mechanical Turk in object detection in images, image classification, sentiment detection, opinion collection, rating, reviewing, and others. In this section we highlight some of the studies utilizing Mechanical Turk in the processes of data collection, or evaluation.

Conducting user studies has ranged between informal surveys to controlled-environment laboratory studies. In such cases, there are tradeoffs between the sample size and the time/monetary cost. Kitter et al. demonstrated that using

Mechanical Turk could be utilized in collecting user-based data points in a cost effective way in regards to time and money [194]. They also warned that during formulating those human intelligence tasks (HITs), a special care is crucial to fully utilize this approach.

In regards to collecting data, we are in need of a large data set that captures the human temporal intention. To do this, prior and during the phases of experimental design, we examined several publications depicting crowd sourcing [195] and most specifically Amazon's Mechanical Turk [196]. Lee and Hu proved that Mechanical Turk could be utilized in generating ground truth data for a similar-scoped study in detecting music moods [161]. In the experimental design, Kosara and Ziemkiewicz conducted several perception and cognition studies on Amazon Mechanical Turk to avoid the problems resulting from poorly designed user studies [197]. Mechanical Turk is also utilized the visualizations field in accessing visualizations design [198].

While searching for which vertical search engines are relevant, Zhou et al. conducted an experiment to prove that relevant verticals derived from different assumptions do correlate with each other [199]. To accomplish this a total of more than 20,000 assessments on 44 search tasks across 11 verticals are collected through Amazon Mechanical Turk and subsequently analyzed.

Delving deeper into the process of completing a human computation task on Mechanical Turk, Heymann and Garcia-Molina developed Turkalytics, which is a tool for gathering data about the workers (or turkers as named henceforth) during the tasks [200]. While Wang et al. analyzed what the recommendations are that could be made to the practitioner to take full advantage of crowdsourcing in general – and Mechanical Turk specifically – and the form of annotation application would best serve the task [201].

CHAPTER 4

LOSS AND PERSISTENCE OF SHARED CONTENT IN SOCIAL MEDIA

"Enlightenment is not an attainment, it is a realization. And when you wake up, everything changes and nothing changes." — Dan Millman, *Way of the Peaceful Warrior*

Based on our review of the various aspects of intention in the context of the social web, human behavioral analysis, crowdsourcing, and shared resource analysis, we decompose the problem into three major components where we will focus our analysis: the shared resource, the concept of time, and the user's behavioral analysis. In this chapter, and as shown in Figure 15, we target the first component of the problem by analyzing the shared resources in social media, their persistence, loss, change, and possibilities of replacement and recovery.
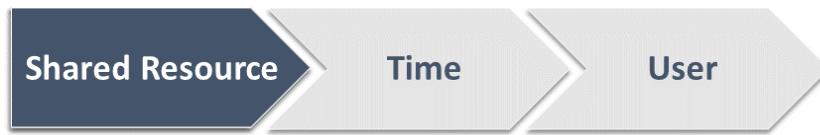
Figure 15. First analysis component: The shared resource in social media

4.1 ESTIMATING SOCIAL MEDIA CONTENT LOSS

Firstly, we analyze the content shared on social networks in an attempt to answer the questions: How much of the social content shared in social networks has been lost [202], and how much can be restored from archives or replaced by similar content [203]?, and is there a relation between the content loss and time [204]?

4.1.1 DATA GATHERING

We compiled a list of URIs that were shared in social media and correspond to specific culturally important events. In this section we describe the data acquisition and sampling process we performed to extract six different datasets which will be tested and analyzed in the following sections.

Stanford SNAP Project Dataset

The Stanford Large Network Dataset is a collection of about 50 large network datasets having millions of nodes, edges, and tuples. It was collected as a part of the Stanford Network Analysis Platform (SNAP) project. It includes social networks, web graphs, road networks, Internet networks, citation networks, collaboration networks, and communication networks. For the purpose of our investigation, we selected their Twitter posts dataset. This dataset was collected from June 1st, 2009 to December 31st, 2009 and contains nearly 476 million tweets posted by nearly 17 million users. The dataset is estimated to cover 20%-30% of all posts published on Twitter during that time frame [205]. To select which events will be covered in this study, we examined CNN's 2009 events timeline[1]. We wanted to select a small number of events that were diverse, with limited overlap, and relatively important to a large number of people. Given that, we selected four events: the H1N1 virus outbreak, the Iranian protests and elections, Michael Jackson's death, and Barrack Obama's Nobel Peace Prize award.

Preparation: Figure 16 shows an example of a tweet record in the SNAP dataset. The record contains the tweet text, the author who posted the tweet, and the timestamp of the tweet. Unfortunately, other useful information is missing, like the original URL of the tweet, the tweet ID, the number of retweets, and the current status of the tweet.

Tag Expansion: We wanted to select tweets that we can say with high confidence are about a selected event. In this case, precision is more important than recall, as collecting every single tweet published about a certain event is less important than making sure that the selected tweets are definitely about that event. Several studies focused on estimating the aboutness of a certain web page or a resource

---

[1]http://www.cnn.com/2009/US/12/16/year.timeline/index.html

Event	Initial Hashtags	Top Co-occurring Hashtags Sample
H1N1 Outbreak	h1n1 (61,351)	swine (61,829)
		swineflu (56,419)
		flu (8,436)
		pandemic (6,839)
		influenza (1,725)
		grippe (1,559)
		tamiflu (331)
MJ Death	michaeljackson (22,934)	michael (27,075)
		mj (18,584)
		thisisit (8,770)
		rip (3,559)
		jacko (3,325)
		kingofpop (2,888)
		jackson (2,559)
		thriller (1,357)
		thankyoumichael (1,050)
Iran Election	iranelection (911,808)	iran (949,641)
		gr88 (197,113)
		neda (191,067)
		tehran (109,006)
		mousavi (16,587)
		freeiran (13,378)
		united4iran (9,198)
		iranrevolution (7,295)
Obama's Nobel	obama (48,161)	peace (3,721)
		nobel (2,261)
		barack (1,292)
		nobelpeace (113)
		nobelpeaceprize (107)

Table 1. Twitter hashtags generated for filtering and their frequency of occurring

in general [97, 100]. Fortunately in Twitter, hashtags incorporated within a tweet can help us estimate their "*aboutness*" as described earlier in Section 3.3. Users normally add certain hashtags to their tweets to ease the search and discoverability in following a certain topic. These hashtags will be utilized in the event-centric filtering process.

For each event, we selected initial tags that describe it (Table 1). Those initial tags were derived empirically after examining some event-related tweets. Next we extracted all the hashtags that co-occurred with our initial set of hashtags, as shown
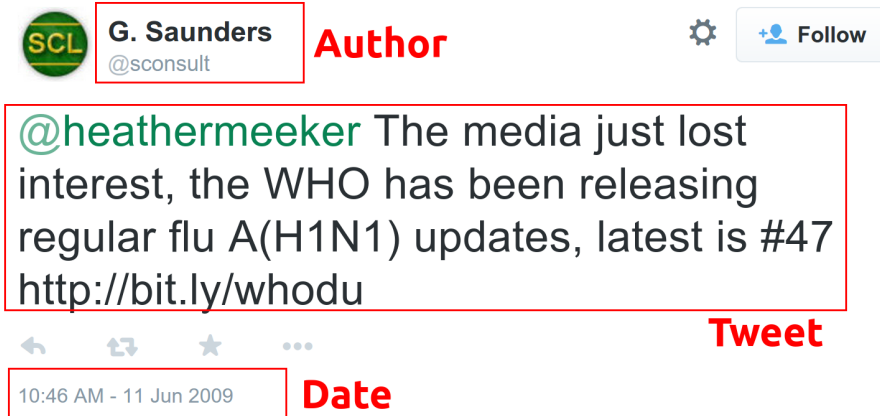
Figure 16. An example of a tweet in SNAP dataset which illustrates typical tweet anatomy

in Figure 16. For example, in class H1N1 we extracted all the other hashtags that appeared along with *#h1n1* within the same tweet and kept count of their frequency. Those extracted hashtags were sorted in descending order of the frequency of their appearance in tweets. We removed all the general scope tags like *#cnn*, *#health*, *#death*, *#war* and others. In regards to aboutness, removing general tags will decrease recall but will increase precision. Finally we picked the top 8-10 hashtags to represent this event-class and are utilized in the filtering process. Table 1 shows the final set of tags selected for each class.

Tweet filtering: In the previous step we extracted the tags that will help us classify and filter tweets in the dataset according to each event. This filtering process aims to extract a reasonable sized dataset of tweets for each event and to minimize the inter-event overlap. Since the life and persistence of the tweet itself is not the focus of this study but rather the associated resource that appears in the tweet (image, video, shortened URI or other embedded resource), we will extract only the tweets that contain an embedded resource. This step resulted in 181 million tweets with embedded resources (i.e., a URI as in Figure 16). These tweets were further filtered to keep only the tweets that have at least one of the expanded tags obtained from Table 1. The number of tweets after this phase reached 1.1 million tweets.

Filtering the tweets based on the occurrence of only one of the hashtags is undesirable as it will cause two problems. First, it will introduce possible event overlap

due to general tweets talking about two or more topics. Second, using only the single occurrence of these tags will yield a huge number of tweets and we need to reduce this size to reach a more manageable size. Intuitively speaking, strongly related hashtags will co-occur often. For example, a tweet that has *#h1n1* along with *#swineflu* and *#pandemic* is most likely about the H1N1 outbreak rather than a tweet having just the tag *#flu* or just *#sick*. Filtering with this co-occurrence will in turn solve both problems: by increasing relevance to a particular event, general tweets that talk about several events will be filtered out thus diminishing the overlap, and in turn it will reduce the size of the dataset.

Next, we increase the precision of the tweets associated with each event from the set of 1.1 million tweets. In the first iteration we selected the tag with the highest frequency of co-occurrence in the dataset with the initial tag and added it to a set we will call the selection set. After that we checked the co-occurrence of all the remaining extracted tags with the tag in the selection set and recorded the frequencies of co-occurrence. After sorting the frequencies of co-occurrence with the tag from the selection set, we picked the highest one to keep and added it to the selection set. We repeated this step of counting co-occurrences but with all the previously extracted hashtags in the selection set from previous iterations.

To elaborate, for H1N1 we assumed that the hashtag *#h1n1* had the highest frequency of appearance in the dataset so we added it to the selection set. In the next iteration we recorded the how many times each tag in the list appeared along with *#h1n1* in a same tweet. If we selected *#swine* as the one with the highest frequency of occurrence with the initial tag *#h1n1* we added it to the selection list and in the next iteration we recorded the frequency of occurrence of the remaining hashtags with both of the extracted tags *#h1n1* and *#swine*. We repeated this step, for each event, to the point where we had a manageable sized dataset with which we were confident in its 'aboutness' in relation to the event.

Two problems appeared from this approach with the Iran and Michael Jackson datasets. In the Iran dataset the number of tweets was in the hundreds of thousands, and even with five tags co-occurrence it was still about 34K+ tweets. To solve this we performed a random sampling from those resulting tweets to take only 10% of

them. The problem with the Michael Jackson dataset was that, upon using five tags to decrease it to a manageable size, we realized there were few unique domains for the embedded resources. A closer look revealed this combination of tags was mostly border-line tweet spam (MJ ringtones). To solve this we used only the two top tags *#michael* and *#michaeljackson*, and then we randomly sampled 10% of the resulting tweets to reach the desired dataset size (Table 2).

Event	Hashtags Selected	Tweets Extracted	Final Tweets
H1N1 Outbreak	h1n1	61,351	**5,517**
	h1n1 & swine	44,972	
	h1n1 & swine & swineflu	42,574	
	h1n1 & swine & swineflu & pandemic	**5,517**	
MJ Death	michael	27,075	**2,293** (10% Sample)
	michael & michaeljackson	**22,934**	
Iran Elections	iran	949,641	**3,429** (10% Sample)
	iran & iranelection	911,808	
	iran & iranelection & gr88	189,757	
	iran & iranelection & gr88 & neda	91,815	
	iran & iranelection & gr88 & neda & tehran	**34,294**	
Obama's Nobel	obama	48,161	**1,118**
	obama & nobel	**1,118**	

Table 2. Tweet filtering iterations and final tweet collections

Egyptian Revolution Dataset

The one year anniversary of the Egyptian revolution was the original motivation to quantify how many resources that were shared during the revolution have persisted during this year [202]. In this case, we started with an event and then tried to get social media content describing it. Despite its ubiquity, gathering social media for a past event is surprisingly hard. We picked the Egyptian revolution due to the role of the social media in curating and driving the incidents that led to the resignation of the president. Several initiatives were commenced to collect and curate the social media content during the revolution like R-sheif.org[2] which specializes in

---

[2]http://www.r-shief.org/

social content analysis of the issues in the Arab world by using aggregate data from Twitter and the Web. Meanwhile, we decided to build our own dataset manually.

There are several sites that curate resources about the Egyptian Revolution and we wanted to investigate as many of them as possible. At the same time, we needed to diversify our resources and the types of digital artifacts that are embedded in them. Tweets, videos, images, embedded links, entire web pages and books were included in our investigation. For the sake of consistency, we limited our analysis to resources created within the period from the 20th of January, 2011 to the 1st of March, 2011. In the next subsections we explain each of the resources we utilized in our data acquisition in detail.

Storify: Storify is a website that enables users to create stories by creating collections of URIs (e.g., Tweets, images, videos, links) and arrange them temporally. These entries are posted by reference to their host websites. Thus, adding content to Storify does not necessarily mean it is archived. If a user added a video from YouTube and after a while the publisher of that video decided to remove it from YouTube the user is left with a gap in their Storify entry. For this purpose we gathered all the Storify entries that were created between 20th of January 2011 and the 1st of March 2011, resulting in 219 unique resources.

IAmJan25: Some entire websites were dedicated as a collection hub of media to curate the revolution. Based on public contributions, those websites collect different types of media, classify them, order them chronologically and publish them to the public. We picked a website, IAmJan25.com, as an example of these websites to analyze and investigate. The administrators of the website received selected videos and images for notable events and actions that happened during the revolution. Those images and videos were selected by users as they vouched for them to be of some importance and they send the resource's URI to the web site administrators. The website itself is divided into two collections: a video collection and an image collection. The video collection had 2387 unique URIs while the image collection had 3525 unique URIs.

Tweets From Tahrir: Several books were published in 2011 documenting the revolution and the Arab Spring. To bridge the gap between books and digital media we analyzed the book ***Tweets from Tahrir*** [206] which was published on April 21st, 2011. As the name states, this book tells a story formed by tweets of people

during the revolution and the clashes with the past regime. We analyzed this book as a collection of tweets that had the luxury of a paperback preservation and focused on the tweeted media, in this case images. The book had a total of 1118 tweets having 23 unique images.

Syria Dataset This dataset was created to represent a current (as of March 2012) event. Using the Twitter search API, we followed the same pattern of data acquisition as in Section 4.1.1. We started with one hashtag, #Syria, and expanded it. Table 3 shows the tags produced from the tag expansion step. After that each of those tags were input into a process utilizing the Twitter streaming API and produced the first 1,000 results matching each tag. From this set, we randomly sampled 10%. As a result, 1955 tweets were extracted, each having one or more embedded resources and tags from the expanded tags in Table 3.

Initial Hashtag	Expanded Hashtags
'Syria'	'Bashar' 'RiseDamascus' 'GenocideInSyria' 'Assad' 'STOPASSAD2012' 'AssadCrimes'

Table 3. Twitter #tags generated for filtering the Syrian Uprising

Table 4 shows the resources collected along with the highest occurring domain names that those resources belong to for each event.

4.1.2 UNIQUENESS AND EXISTENCE

From the previous data gathering step we obtained six different datasets related to six different historic events. For each event we extracted a list of URIs that were shared in tweets or uploaded to sites like Storify or IAmJan25. To answer the question of how much of the social media content is missing, we tested those URIs for each dataset to eliminate URI aliases in which several URIs identify the same resource. Upon obtaining those unique URIs we examine how many are still available on the live web as shown in Figure 17. We also calculate how many are available in public web archives.

Event	Top Domain Names	Resources Found
MJ	youtube	110
	twitpic	45
	latimes	43
	cnn	30
Iran	youtube	385
	twitpic	36
	blogspot	30
	roozonline	29
H1N1	rhizalabs	676
	reuters	17
	google	16
	flutrackers	16
Obama	blogspot	16
	nytimes	15
	wordpress	12
	youtube	11
Egypt	youtube	2,414
	cloudfront	2,303
	yfrog	1,255
	twitpic	114
Syria	youtube	130
	twitter	61
	hostpic.biz	9
	telegraph.co.uk	5

Table 4. The top level domains found for each event ordered descendingly by the number of resources

Figure 17. Analysis of how much of the shared content is still on the live web

Uniqueness

Some URIs, especially those that appear in Twitter, may be aliases for the same resource. For example "http://bit.ly/2EEjBl" and "http://goo.gl/2ViC" both resolve to "http://www.cnn.com". To solve this, we resolved all the URIs following redirects to the final URI. The HTTP response of the last redirect has a field called *location* that contains the original long URI of the resource. This step reduced the total number of URIs in the six datasets from 21,625 to 11,051. Table 5 shows the number of unique resources in every dataset.

Existence on the Live Web

After obtaining the unique URIs from the previous step we resolve all of them and classify them as Success or Failure. The *Success* class includes all the resources that ultimately return a "200 OK" HTTP response. The *Failure* class includes all the resources that return a "4XX" family response like "404 Not Found", "403 Forbidden" and "410 Gone"; the "30X" redirect family while having infinite loop redirects; and server errors with response "50X". To avoid transient errors, we repeated the requests, on all datasets, several times for a week to resolve those errors.

We also tested for "Soft 404s", which are pages that return "200 OK" response code but are not a representation of the resource, using a technique based on a heuristic for automatically discovering soft 404s from Bar-Yossef et al., as shown in Section 3.3.2 [78]. We also include no response from the server, as well as DNS time-outs, as failures. Note that failure means that this resource is **missing** on the live

	All	Unique	
	5,517	1,645=**29.82%**	
_H1N1_	Archived	Not Archived	
Available	595=36.17%	656=39.88%	
Missing	98=5.96%	296=17.99%	394=**23.95%**
	693=**42.12%**	each/1,645	

	All	Unique	
	2,293	1,187=**51.77%**	
_MJ_	Archived	Not Archived	
Available	316=26.62%	474=39.93%	
Missing	90=7.58%	307=25.86%	397=**33.45%**
	406=**34.20%**	each/1,187	

	All	Unique	
	3,429	1,340=**39.08%**	
Iran	Archived	Not Archived	
Available	415=30.97%	586=43.73%	
Missing	101=7.54%	238=17.76%	339=**25.30%**
	516=**38.51%**	each/1,340	

	All	Unique	
	1,118	370=**33.09%**	
Obama	Archived	Not Archived	
Available	143=38.65%	135=36.49%	
Missing	33=8.92%	59=15.95%	92=**24.86%**
	176=**47.57%**	each/370	

	All	Unique	
	7,313	6,154=**84.15%**	
Egypt	Archived	Not Archived	
Available	1,069=17.37%	4440=72.15%	
Missing	173=2.81%	472=7.67%	645=**10.48%**
	1242=**20.18%**	each/6,154	

	All	Unique	
	1,955	355=**18.16%**	
Syria	Archived	Not Archived	
Available	19=5.35%	311=87.61%	
Missing	0=0%	25=7.04%	25=**7.04%**
	19=**5.35%**	each/355	

Table 5. Percentages of unique resources for each event and the percentages of presence of those unique resources on live web and in archives. All resources = 21,625, unique resources = 11,051

web. Table 5 summarizes, for each dataset, the total percentages of the resources missing from the live web and the number of missing resources divided by the total number of unique resources.

Existence in the Archives

In the previous step we tested the existence of the unique list of URIs for each event on the live web. Next, we evaluated how many URIs have been archived in public web archives. To check those archives we utilized the Memento framework, as described in Section 2.4. If there is a memento for the URI, we downloaded its memento TimeMap and analyzed it. The TimeMap is a datestamp ordered list of all known archived versions (or "mementos") of a URI. Next, we parsed this TimeMap and extracted the number of mementos that point to versions of the resource in the public archives. We declared the resource to be archived if it has at least one memento. This step was also repeated several times to avoid transient errors in the archives before deeming a resource as unarchived. The results of this experiment along with the archive coverage percentage are also presented in Table 5.

4.1.3 EXISTENCE AS A FUNCTION OF TIME

Inspecting the results from the previous steps suggests that the number of missing shared resources in social media corresponding to an event is directly proportional to age. To determine dates for each of the events this we extracted all the creation dates from all the tweet-based datasets and sorted them. For each event, we plotted a graph illustrating the number of tweets per day related to that event as shown in Figure 18. Since the dataset is separated temporally into three partitions, and in order to display all the events on one graph we reduced the size of the x-axis by removing the time periods not covered in our study.

Upon examining the graph we found an interesting phenomena in the non-Syrian and non-Egyptian events: each event has two peaks. Upon investigating history timelines we came to the conclusion that those peaks reflect a second wave of social media interaction as a result of new incident within the same event after a period of time. For example, in the H1N1 dataset the first peak illustrates the world-wide outbreak announcement, while the second peak denotes the release of the vaccine. In the Iran dataset, the first peak shows the peak of the elections while the second
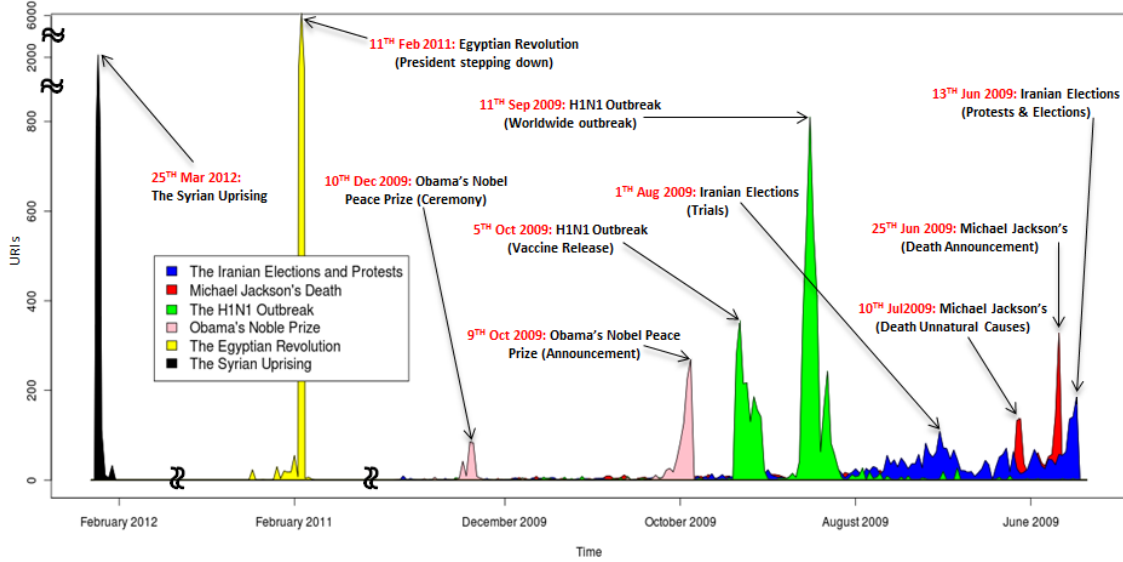
Figure 18. URIs shared per day corresponding to each event and showing the two peaks in the non-Syrian and non-Egyptian events. Note: the x-axis has two time breaks and it flows from the present to the past

peak pinpoints the Iranian trials. As for the MJ dataset the first peak corresponds to his death and the second peak describes the rumors that Michael Jackson died of unnatural causes and a possible homicide. For the Obama dataset, the first peak reveals the announcement of his winning the prize while the second peak presents the award-giving ceremony in Oslo. For the Egyptian evolution, the resources are all within a small time slot of two weeks around the date 11th of February. As for the Syrian event, since the collection was very recent, there was no obvious peaks. Those peaks we examined will become temporal centroids of the social content collections (the datasets): MJ (June 25th & July 10th, 2009), Iran (June 13th & August 1st, 2009), H1N1 (September 11th & October 5th, 2009), and Obama (October 9th & December 10th, 2009). Egypt was February 11th, 2011, and the Syria dataset also had one centroid on March 27th, 2012. We split each event according to the two centroids in each event accordingly. Figure 18 shows those peaks and Table 6 shows the missing content and the archived content percentages corresponding to each centroid.

Figure 19 shows the missing and archived values from Table 6 as a function of time since shared. Equation 1 shows the modeled estimate for the percentage of

	MJ		Iran		Egypt
Missing %	36.24%	31.62%	26.98%	24.47%	10.48%
Archived %	39.45%	30.78%	43.08%	36.26%	20.18%

	Obama		H1N1		Syria
Missing %	24.59%	26.15%	23.49%	25.64%	7.04%
Archived %	47.87%	46.15%	41.65%	43.87%	5.35%

Table 6. The split dataset
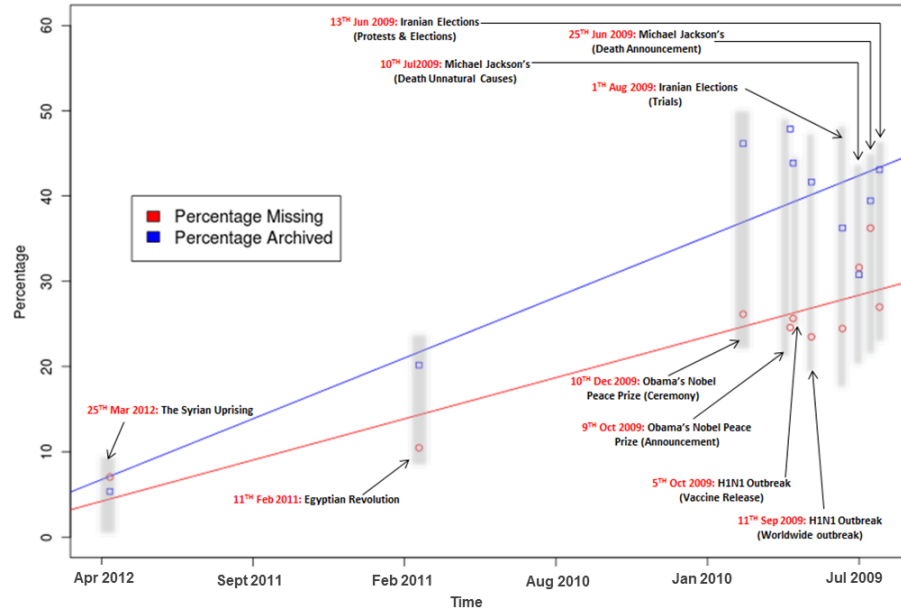
Figure 19. Percentage of content missing and archived for the events as a function of time. The gray bars are present solely for visual alignment

shared resources lost, where *Age* is in days. While there is a less linear relationship between time and being archived, Equation 2 shows the modeled estimate for the percentage of shared resources archived in a public archive.

$$Content\ Lost\ Percentage = 0.02(Age\ in\ days) + 4.20 \tag{1}$$

$$Content\ Archived\ Percentage = 0.04(Age\ in\ days) + 6.74 \tag{2}$$

Given these observations and our curve fitting, we estimate that after a year from publishing about 11% of content shared in social media will be gone. After

this point, we are losing roughly 0.02% of this content per day.

We can conclude that there is a nearly linear relationship between time of sharing in the social media and the percentage lost. Although not as linear, there is a similar relationship between the time of sharing and the expected percentage of coverage in the archives. To reach this conclusion, we extracted collections of tweets and other social media content that was posted and shared in relation to six different events that occurred in the time period from June 2009 to March 2012. Next we extracted the embedded resources within this social media content and tested their existence on the live web and in the archives. After analyzing the percentages lost and archived in relation to time and plotting them we used a linear regression model to fit those points. Finally, we presented two linear models that can estimate the existence of a resource, that was posted or shared at one point of time in the social media, on the live web and in the archives as a function of age in the social media. The next step is to validate this modeling and analyze the uniformity of the predicted disappearance of resources. Furthermore, we investigate methods to deal with this loss of resources by providing viable replacements.
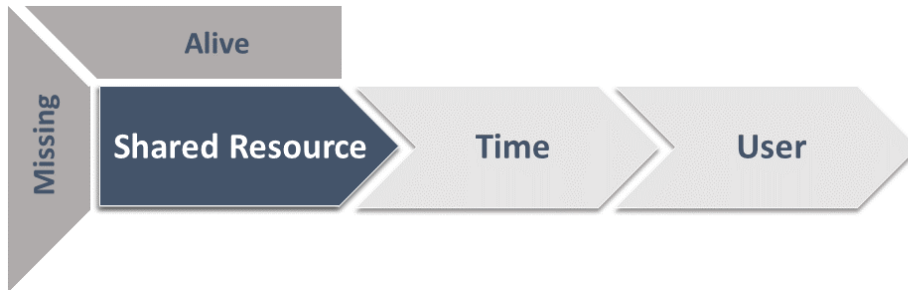
Figure 20. Analysis of how much of the shared content is missing and stays missing

4.2 PERSISTENCE AND STABILITY OF SHARED RESOURCES

A year after building the predictive model of resource existence and archival elaborated in the previous section, we decided to revisit our model and investigate if the relationship with time still holds or not. This validation will provide a better understanding of the persistence and the stability of loss across time (as shown in Figure 20), and pave the way towards overcoming this loss in resources. On the same dataset we reran the experiment and discovered a phenomenon of reappearance and
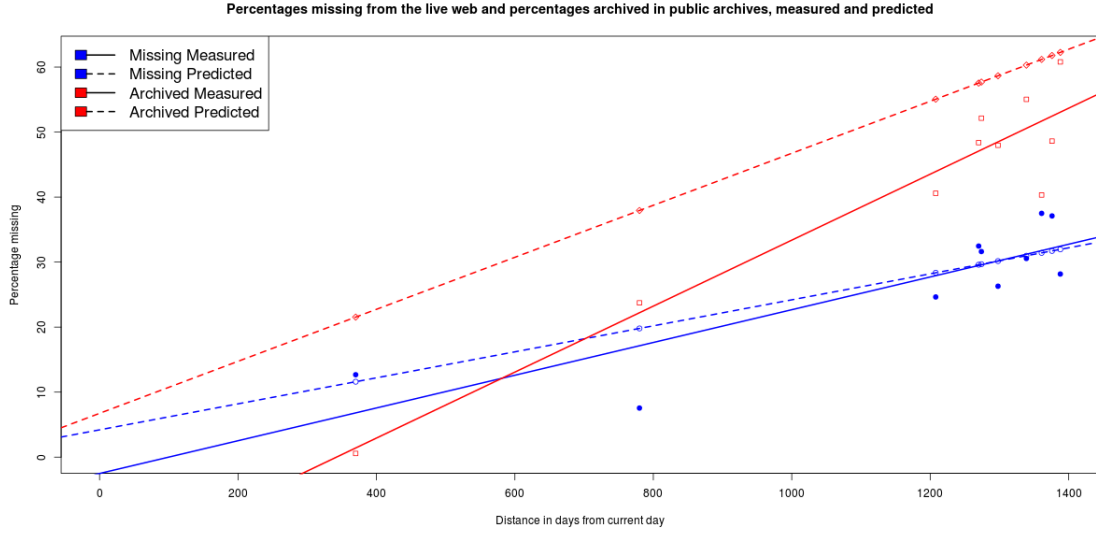
Figure 21. Measured and predicted percentages of resources missing and archived
for each dataset and the corresponding linear regression

disappearance that was interesting to report [203].

4.2.1 REVISITING EXISTENCE

In the model estimated in our previous experiment in 2012, we found a nearly linear relationship between the amount missing from the web and time as shown earlier in Equation 1. We also found a less linear relationship between the amount archived and time as shown in Equation 2.

After a year had passed, we wanted to analyze our findings and the estimation calculated to see if it still matches our prediction. For each of the six datasets investigated, we repeated the same experiment of analyzing the existence of each of the resources on the live web. A resource is deemed missing if it returned an HTTP response other than 200 OK. A resource is considered missing as well if it was declared a "soft 404".

Table 7 shows the results from repeating the experiment, the predicted calculated values based on our model, and the corresponding errors. Figure 21 illustrates the measured and the estimated plots for the missing resources. The standard error

Missing			
	Measured	**Predicted**	**Error**
MJ	37.10%	31.72%	5.38%
Death	37.50%	31.42%	6.08%
Iran	28.17%	31.96%	3.79%
Elections	30.56%	30.98%	0.42%
H1N1	26.29%	30.16%	3.87%
Outbreak	31.62%	29.68%	1.94%
Obama's	32.47%	29.60%	2.87%
Nobel	24.64%	28.36%	3.72%
Egypt	7.55%	19.80%	12.25%
Syria	12.68%	11.54%	1.14%
Average Prediction Error			**4.15%**

Archived			
	Measured	**Predicted**	**Error**
MJ	48.61%	61.78%	13.17%
Death	40.32%	61.18%	20.86%
Iran	60.80%	62.26%	1.46%
Elections	55.04%	60.30%	5.26%
H1N1	47.97%	58.66%	10.69%
Outbreak	52.14%	57.70%	5.56%
Obama's	48.38%	57.54%	9.16%
Nobel	40.58%	55.06%	14.48%
Egypt	23.73%	37.94%	14.21%
Syria	0.56%	21.42%	20.86%
Average Prediction Error			**11.57%**

Table 7. Measured and predicted percentages for missing and archived content in each dataset

calculated is equal to 4.15% which shows that our model still holds and it presents a good realistic prediction.

To verify the second part of our model we calculated the percentages of resources that are archived at least once in one of the public archives. Table 7 illustrates the results measured, predicted, and the corresponding standard error as well. Figure 21 also displays the measured and predicted corresponding plots for the archived resources.

In case of modeling the content missing, we verified that the percentages have a direct relationship with time and our previous prediction model is considerably

accurate, with an average standard error of 4.15%. The archived content percentages had a higher error percentage of 11.57% and became less linear with time. This fluctuation in the archival percentages convinced us that further analysis is needed.

4.2.2 REAPPEARANCE AND DISAPPEARANCE

In measuring the percentage of resources missing from the live web, we assumed that when a resource is deemed to be missing it remains missing. We also assumed that if a snapshot of the resource is present in one of the public archives the resource is deemed to be archived and that this snapshot persists indefinitely. Utilizing the response logs resulting from the existence experiment in 2012 and in 2013, we compare the corresponding HTTP responses and the number of mementos for each resource. As expected, portions of the datasets disappeared from the live web and were labeled as missing. An interesting phenomena occurred as several of the resources that were previously declared as missing became available on the live web as shown in Table 8.

A possible explanation of this reappearance could be a domain or a webserver being disrupted and restored again. For example, the 1000memories.com site was down in 2012 but was eventually restored [202]. Another possible explanation is that the previously missing resources could be linked to a suspended user account that was reinstated. To eliminate the effect of transient errors, the experiment was repeated three times in the course of two weeks. To grasp a better understanding of resource existence we model the probability of reappearance of a resource that was deemed missing. A more accurate notion of existence would be the collective percentage of disappearance and reappearance of a resource at any given time, as explained in Equation 3.

$$Missing = Disappearance - Reappearance \qquad (3)$$

Corresponding to each of the six events, and comparing the responses recorded in 2012 and in 2013, Figure 22 illustrates the percentages of the resources reappearing in the corresponding datasets. Given those percentages we notice a linear relationship with time. By applying linear regression in curve fitting, we reached Equation 4, describing the reappearance of resources as a function of time.

$$LiveContent\ Reappearing = 0.01(Age\ in\ days) - 1.42 \qquad (4)$$

Event	Percentage		
	Re-appearing on the web	Disappearing from archives	Going from 1 memento to 0
MJ	11.29%	9.98%	2.72%
Iran	11.48%	11.17%	2.89%
Obama	6.63%	15.65%	4.24%
H1N1	3.68%	5.46%	1.96%
Egypt	4.21%	2.81%	0.23%
Syria	1.97%	2.25%	0.28%
Average	**6.54%**	**7.89%**	**2.05%**

Table 8. Percentages of resources reappearing on the live web and disappearing from the public archives per event

In the previous experiment, we modeled the archival existence or the percentage archived as a function of time. The phenomena analyzed in the previous section showed the instability of the resources in the web which influenced us to investigate the archived resources as well. We deemed a resource to be archived if there existed at least one publicly available memento of the resource in the archives. For each resource we extracted the memento TimeMaps and recorded the number of available mementos. The resources are expected to have the same number of mementos or more, indicating more snapshots taken into the archives or unarchived resources started to exist in the archives. We noticed another interesting phenomena: the number of available mementos of several resources have actually decreased, indicating disappearance from the archives as shown in Table 8. A possible explanation could be due to TimeMap shrinkage, as in past revisions of the Memento aggregator, search engine caches were represented as archives. Brunelle and Nelson explained that the number of mementos in a TimeMap in some scenarios would decrease: for example, archival redaction of some or all of the mementos, archival restructuring, and transient errors of one or more archives [207]. In the recent revision, search engine caches are no longer used as archives, which we estimate by measuring the number of resources whose TimeMaps went from one memento to zero as shown in Table 8. Similarly, we plot the percentages of memento disappearance in Figure 22. Equation 5 results from applying linear regression in curve fitting. Inspecting Figure 22 verifies to a certain degree our explanation of the archival disappearance phenomena as the regression line maintains the same slope of the estimated model
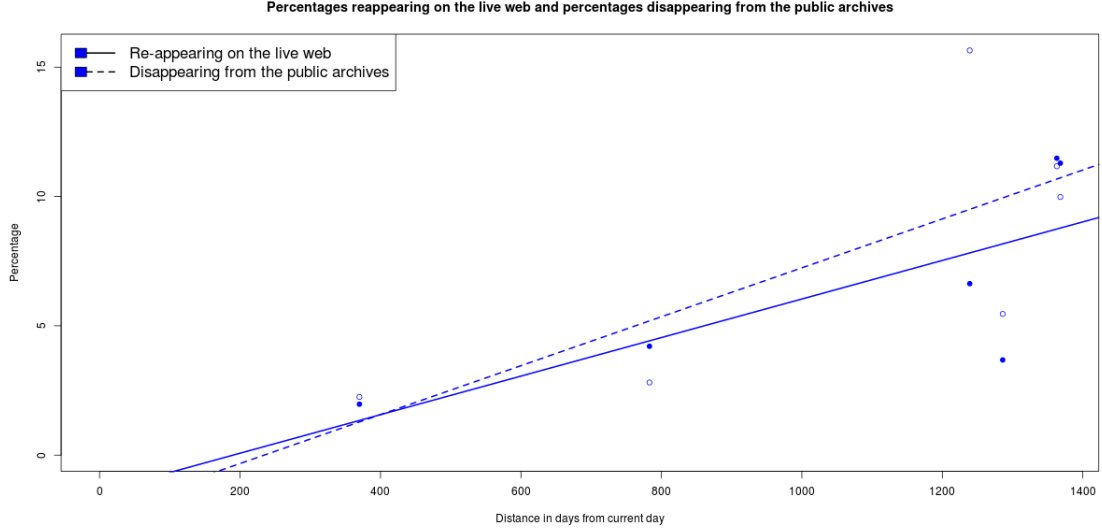
Figure 22. Percentages of resources reappearing on the live web and the resources disappearing from the public archives

as shown in Figure 21, while it differs in the Y-intercept.

$$Mementos\ Disappearing = 0.01(Age\ in\ days) - 2.22 \tag{5}$$

4.2.3 TWEET EXISTENCE

After focusing on the embedded resources shared in posts in social media another question arose: what about the existence of the social post itself? In collecting the dataset that we utilized in our analysis we focused on the embedded resource and the creation dates. Also, the SNAP dataset we used provides only the tweet text, the author's username, and the creation date with no further information about the tweet or its URI. A social post could face the same fate of the embedded resource by being deleted, service hosting it discontinued, or the author's account getting suspended. Similarly to the resource existence testing, we checked the existence of the posts by examining the HTTP response headers. To work around the absence of the tweet URI, we utilized Topsy, a service that mines social media websites like Twitter to provide analytics and insight to topics and resources. Using the API, we can extract all the available tweets that incorporate a given URI with a maximum of 500 tweets. For each resource in the dataset we extract all the tweets and check their

existence on the live web accordingly. Given a URI, we can estimate the percentage of social posts that are missing. This number could give an insight to what is the probability that the post itself went missing. Table 9 shows the results for each dataset. Figure 23 illustrates the collective percentages through time. Equation 6 shows the result of curve fitting the percentages of loss as a function of time.

$$SocialPosts\ Missing = 0.01(Age\ in\ days) + 0.88 \qquad (6)$$

Event	H1N1	MJ	Iran	Obama	Egypt	Syria	Average
Average% of missing posts	14.43%	14.59%	10.03%	7.38%	15.08%	0.53%	10.34%

Table 9. Average percentage of missing posts
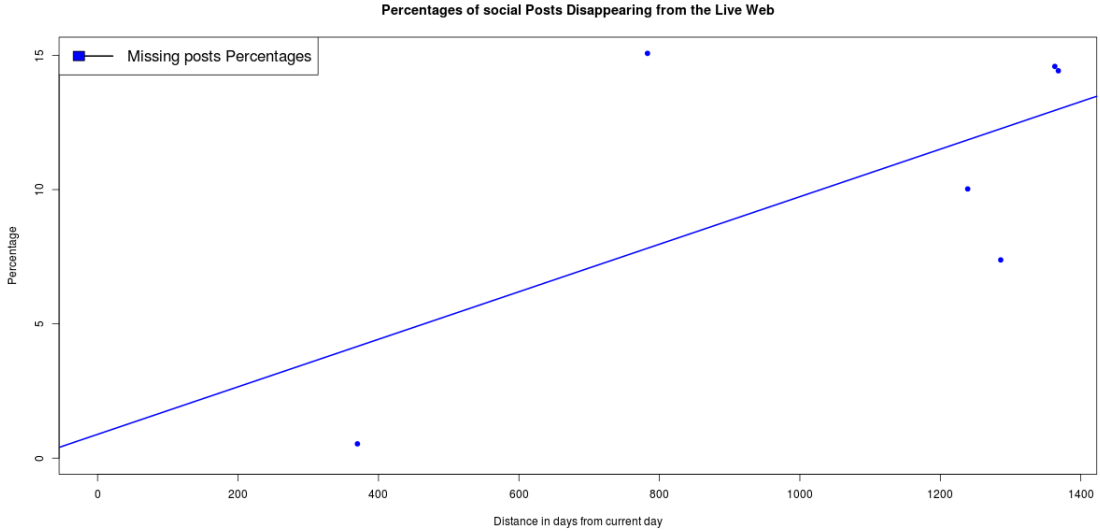
Figure 23. Percentages of missing posts averages curve fitted using linear regression

4.3 RECONSTRUCTING THE MISSING WEB

The evolution of the role of social media and the ease of reader interaction played a crucial part in information dissemination and preservation. We argue that social media could be utilized to discover replacement resources for the unarchived shared resources. To elaborate, when a user tweets about something or creates a Facebook
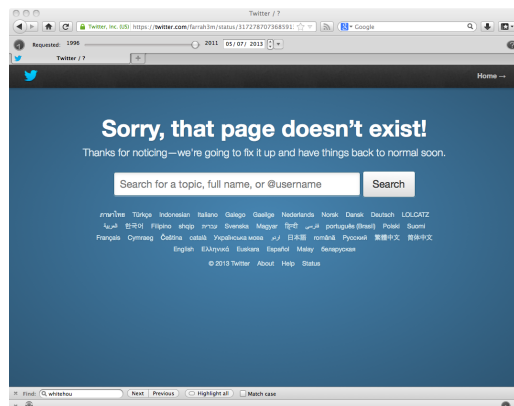
(a) The deleted tweet by user @Farrah3m
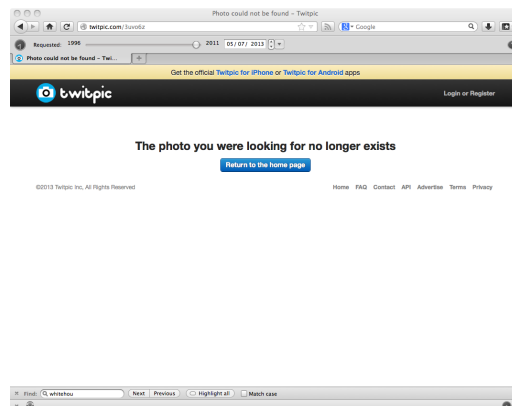
(b) The corresponding image attached to the deleted tweet by @Farrah3m

(c) The Topsy page corresponding to the deleted twitpic image and tweet

(d) The high resolution image replacement to the deleted TwitPic[a]

——————————————

[a]http://gdb.voanews.com/
703A8C3D-DC13-40E1-95B1-
F5688642D2AA_cx0_cy7_cw0_mw1024_s_n_r1.jpg

Figure 24. Tweet image replacement example

post, it leaves behind a trail of copies, links, likes, comments, other shares. If the shared resource is later gone, these traces, in most cases, still persist. To elaborate, on January 28, 2011, three days into the fierce protests that would eventually oust the Egyptian President Hosni Mubarak, a Twitter user (@Farrah3m) posted a link to a picture that supposedly showed an armed man as he ran on a "rooftop during clashes between police and protesters in Suez". Since then, the tweet has been deleted https://twitter.com/Farrah3m/status/31727870736859137 as shown in Figure 24a. The image associated with the tweet on the twitpic service has been deleted as well http://twitpic.com/3uvo6z as shown in Figure 24b. The user @Farrah3m still exists, but she has deleted many of her tweets from during the Egyptian Revolution.

But if we prepend the twitpic URI with "topsy.com/" to get: `http://topsy.com/http://twitpic.com/3uvo6z`, we see the original tweet, and a small but not full-size version of the image as shown in Figure 24c. We were able to find the high resolution replacement of this thumbnail size image, which was taken by Reuters photographer Mohamed Abd El-Ghany as shown in Figure 24d.

Thus we wanted to automate the process of replacement discovery and in this experiment we investigated if the other tweets that also linked to the resource can be mined to provide enough context to discover similar resources that can be used as a substitute for the missing resource, as shown in Figure 25. To do this, we extracted up to the 500 most recent tweets about linked URIs and we proposed a method of finding the social link neighborhood of the social post and the resource we are attempting to reconstruct. This link neighborhood could be mined for context identifiers and alternative related resources.
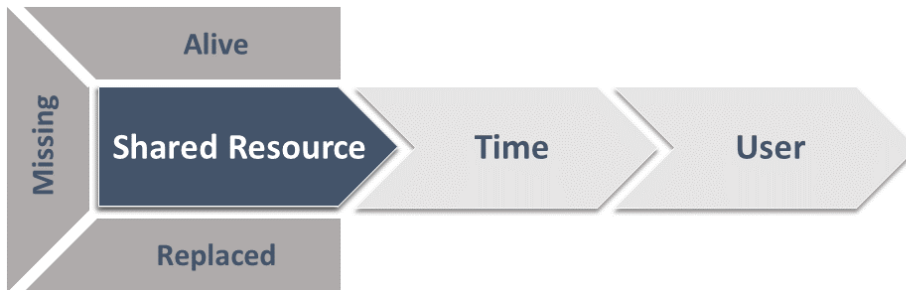
Figure 25. Analyze the possibility of finding replacements/recostructs to the missing content

4.3.1 CONTEXT DISCOVERY AND SHARED RESOURCE REPLACEMENT

A web resource can fall into one of the categories shown below. These categories were adopted from the work of McCown and Nelson [136].

	Archived	**Not Archived**
Available	Replicated	Vulnerable
Missing	Endangered	Unrecoverable

Table 10. Different states of a web resource

If a resource was currently available on the live web and also archived in public archives then it is considered replicated and safe. The resource is considered vulnerable if it persists on the web but has no available archived versions. The vulnerability relies on the fact that the resource is prone to complete loss, as shown in our previous study. If a resource is not available on the live web but has an archived version at least then it is considered endangered, as it relies on the stability and the persistence of the archive. The worst case scenario occurs when the resource disappears from the live web without being archived at all, thusly to be considered unrecoverable. In our study we focus on the latter category and how we can utilize the social media in identifying the context of the shared resource and elect a possible replacement candidate to fill in the position of the missing resource and maintain the same context of the social post.

A shared resource leaves traces, even after it ceases to exist on the web. We attempt to collect those traces and discover context for the missing resource. Since Twitter, for example, restricts the length of the posts to be 140 characters, an author might rely mostly on the shared resource in conveying a thought or an idea by embedding a link in the post and resorting to limiting the associated text. Thus, obtaining context is crucial when the resource disappears. To accomplish that, we tried to find the social link neighborhood of the tweet and the resource we were attempting in this context discovery. When a link is shared on Twitter for example, it could be associated with describing text in the form of the status itself, hashtags, usertags, or other links as well, as shown in Figure 16. These co-existing links could act as a viable replacement to the missing resource under investigation while the tags and text could provide better context enabling a better understanding of the resource.

Social Extraction

Given the URI of the resource under investigation, we utilized Topsy's API to extract all the available tweets incorporating this URI. Fortunately, Topsy's API handles these shortened URIs by searching their index for the final target URI rather than the shortened form. A maximum of 500 tweets of the most recent tweets posted can be extracted from the API regarding a certain URL. The content from all the tweets is collected to form a "social context corpus".

From this corpus, we extract the best replacement tweet by calculating the

longest common N-gram. This represents the tweet with the most information that describes the target resource intended by the author. Within some tweets, multiple links coexist within the same text. These co-occurring resources share the same context and maintain a certain relevancy in most cases. A list of those co-occurring resources are extracted and filtered for redundancies. Finally, the textual components of the tweets in the corpus are extracted after removing usertags, URIs, social interaction symbols like "RT". We named the document composed of those text-only tweets in the form of phrases the "Tweet Document".

Figure 26 illustrates the JSON object produced from analyzing the extracted social context corpus of a resource, as described above.

```
Reconstruction:
{
  "URI": "http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-
  year.html",
  "Related Tweet Count": 290,
  "Related Hashtags": "#history #jan25 #sschat #arabspring #jrn112 #archives
    #in #revolution #iipc12 #mppdigital #egypt #recordkeeping #twitter
    #egyptrevolution #digitalpreservation #preservation #webarchiving #or2012
    #1anpa #socialmedia",
  "Users who talked about this": "@textfiles @jigarmehta @blakehounshell]
    @jonathanglick @daensen404: @ryersonjourn @chanders @theotypes) @jwax55
    @marklittlenews @ndiipp ...",
  "All associated unique links:": "http://t.co/ZRASTg5o http://t.co/eXhlSTRF
    http://t.co/3GIb6oI3 http://t.co/ArVqCqfP ...",
  "All other links associated:": "http://www.cs.odu.edu/~mln/pubs/
    tpdl-2012/tpdl-2012.pdf http://dashes.com/anil/2011/01/if-you
    -didnt-blog-it-it-didnt-happen.html",
  "Most frequent link appearing:": "http://t.co/0A1q2fzz",
  "Number of times the Most frequent link appearing:": 19,
  "Most frequent tweet posted and reposted:": "@acarvin You may have seen this
    already. Arab Spring digital content is apparently being lost.",
  "Number of times the Most frequent tweet appearing:": 23,
  "The longest common phrase appearing:": "You may have seen this already Arab
    Spring digital content is apparently being lost",
  "Number of times the Most common phrase appearing:": 28
}
```

Figure 26. JSON object produced from analyzing a resource's extracted *social context corpus* using the Topsy API

Resource Replacement Recommendation

From the social extraction phase above we gathered information that helps us to infer the aboutness and context of a resource. Given this context, can we utilize it in obtaining a viable replacement resource to fill in the missing one and provide the same context?

To answer this, we utilize the work of Klein et al. [128] in defining the lexical signatures of web pages, as discussed in Section 3.3.1. First, we extract the tweet document as described above. Next, we removed all the stop words and applied Porter's stemmer to the remaining words[3]. We calculated the term frequency of each stemmed word and sorted them from highest to lowest occurrence. We converted each stem to a corresponding original word. Finally, we extracted the top five words to form our tweet-based lexical signature, or "Tweet Signature".

On the one hand, and using this tweet signature as a query, we utilized Google's search engine to extract the top ten resulting resources. On the other hand, we collected all the other co-occurring pages in the tweets obtained by the API. These pages combined produce a replacement candidate list of resources. One or more of these can be utilized as a viable replacement of the resource under investigation.

To choose which resource is more relevant and a possibly better replacement we utilized once more the tweet document extracted earlier. For each of the extracted pages in the candidate list, we downloaded the representation and utilized the boilerpipe library in extracting the text within, as demonstrated by Kohlschutter et al. [208]. The library provides algorithms to detect and remove the surplus "clutter" (boilerplate, templates) around the main textual content of a web page. Having a list of possible candidate textual documents and the tweet document, the next step was to calculate similarity. We utilized cosine similarity to sort pages according to the measured value of similarity to the tweets' page describing the resource under reconstruction.

At this stage we extracted contextual information about the resource and a possible replacement. The next step was to measure how well the reconstruction process was undergone and how close this replacement page was to the missing resource.

---

[3]https://pypi.python.org/pypi/stemming/1.0

4.3.2 REPLACEMENT EVALUATION

Since we could not measure the quality of the discovered context or the resulting replacement page to the missing resource, we had to set some assumptions. We extracted a dataset of resources that are currently available on the live web and assumed they do not exist anymore. Each of these resources are textually-based and neither media files nor executables. Each of these resources has to have at least 30 retrievable tweets using Topsy's API to be enough to build context.

We collected a dataset of 472 unique resources following these rules. We performed the context extraction and the replacement recommendation phases. We downloaded the resource under investigation ($R_{missing}$) and the list of replacements from the search engines (R_{search}) and the list of co-occurring resources ($R_{co-occurring}$). For each, we used the boilerpipe library to extract text and use cosine similarity to perform the comparisons. For each resource, we measured the similarity between the $R_{missing}$ and the extracted tweet page. For each element in R_{search}, we calculated the cosine similarity with the tweet page and sort the results accordingly from most similar to the least. We repeated the same with the list of co-occurring resources $R_{co-occurring}$. Then we calculated the similarity between $R_{missing}$ and $R_{search}(first)$, indicating the top result obtained from the search engine index. Then, we compared $R_{missing}$ with each of the elements in R_{search} and $R_{co-occurring}$ to demonstrate the best possible similarity.

Figure 27 illustrates the different similarities sorted for each measure. From the graph we can state that 41% of the time, we can extract a significantly similar replacement page $R_{replacement}$ to the original resource $R_{missing}$ ($\geq 70\%$). Finally, the mean reciprocal rank (MRR) = 0.43.

In conclusion, we verified our previous analysis and estimation of the percentage missing of the resources shared on social media in Section 4.2. The content dissapearence function of time described by Equation 1 still holds. As for the model estimated for the amount archived, it showed an alteration. The slope of the regression line in the model stayed the same while the y-intercept varied. We deduce that a possible explanation to this phenomena is due to TimeMap shrinkage. Previously, TimeMaps incorporated search engine caches as mementos, and this is no longer valid. This explains to a certain degree the uniform variation in the estimated function. Unfortunately, we cannot verify this precisely, as we do not have the past TimeMaps.
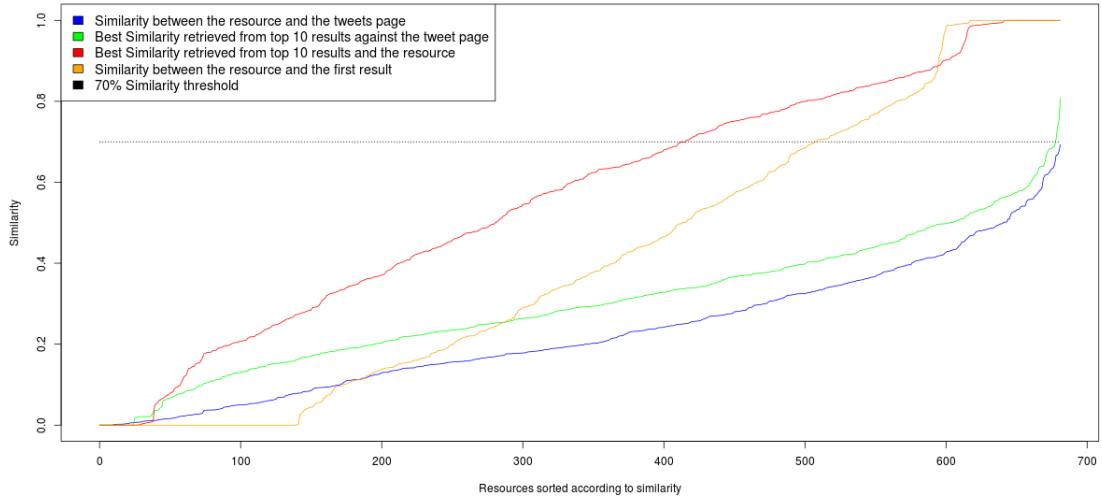
Figure 27. Similarities with the original resource $R_{missing}$

Next, we classified web resources into four different categories in regards to existence on the live web and in public web archives. Then we addressed the unrecoverable category, where the resource is deemed missing from the live web whilst not having any archived versions. Since we could not perform a full reconstruction or retrieval, we utilized the social nature of the shared resources by using Topsy's API in discovering the resource's context. Using this context and the co-occurring resources, we applied a range of heuristics and comparisons to extract the most viable replacement to the missing resource from its social neighborhood.

Finally, we performed an evaluation to measure the quality of this replacement and found that for 41% of the resources, we could obtain a significantly similar replacement resource with $\geq 70\%$ similarity.

CHAPTER 5

FOOTPRINTS IN THE WEB

"Intuition is really a sudden immersion of the soul into
the universal current of life." — Paulo Coelho, *The
Alchemist*

In Chapter 4 we demonstrated that content on the web is susceptible to loss
or change. In this chapter we explore the dimension of time in relation to the re-
sources and the users, as shown in Figure 28. Linking to the previous chapter we
start by analyzing the effect of time on shared content and the experienced change,
possibly affecting the author's initial intention. We showed long-term change (span-
ning months, years) with irregular observations. In this chapter, we start with a
longitudinal study measuring content change, sharing schemes, and the relationship
between them (Section 5.1) with the emphasis to quantify change in the short-term
and with regular observations. Then we explore the past web by analyzing how
much of the web is already archived (Section 5.2). Finally, we illustrate methods of
estimating the age of shared content (Section 5.3).
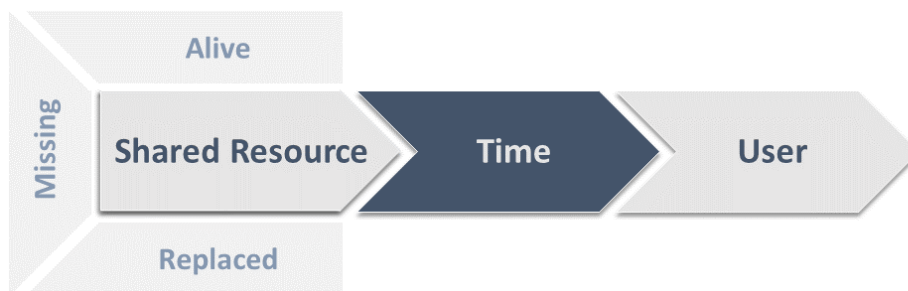
Figure 28. Second analysis component: Time

5.1 MEASURING SHORT-TERM CHANGE IN SHARED RESOURCES

In 2012, we established that resources linked in tweets from six socially important historical events were disappearing ("404 Not Found" response code) at a rate of about 11% for the first year and 7% per year afterwards (Section 4.1). In 2013, we verified that this rate of loss is still holding up (Section 4.2). However, we have not attempted to measure what percentage of the live web resources are off-topic (that is, still "200 OK" but no longer are about the tweet in which they were linked), indicating a shift in the intention through time.
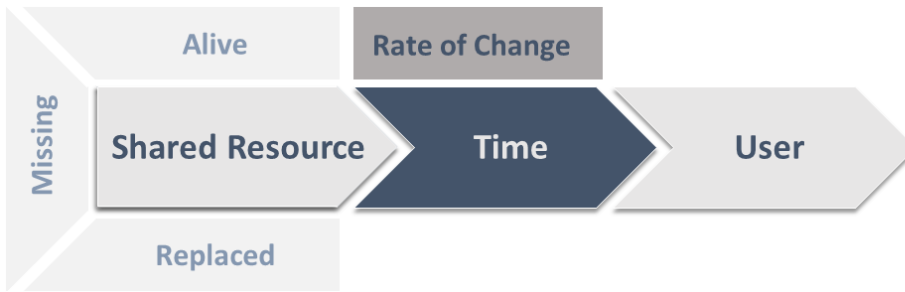
Figure 29. Longitudinal study: Rate of change of shared content

While there has been significant related work about studying the change of web pages (Section 3.3.2), we are interested in a fine-grained study about how much pages change before and after they are linked in social media and how this change affects their dissemination and sharing trends (Figure 29). Even popular sites like cnn.com are archived only a couple of times per day at the Internet Archive; this is too infrequent to detail the changes between t_{tweet} and t_{click}.

To understand this minute, rapid change we started a pilot study for this sole purpose. Using the Twitter public timeline we assembled a list of shortened URIs that were freshly shared on Twitter. We collected these URIs by querying the Twitter API for tweets having a bitly shortened URI. The reason behind this choice is the using the Bitly API we can extract the creation date of the URI or the time it was shortened. Furthermore, we can extract a multitude of useful features like a total click log since the creation date, referring sites and countries, and others.

The first question was: for the content that is being shared now, when was it created? We collected a random sample of 4,000 tweet-bitly pairs from the Twitter
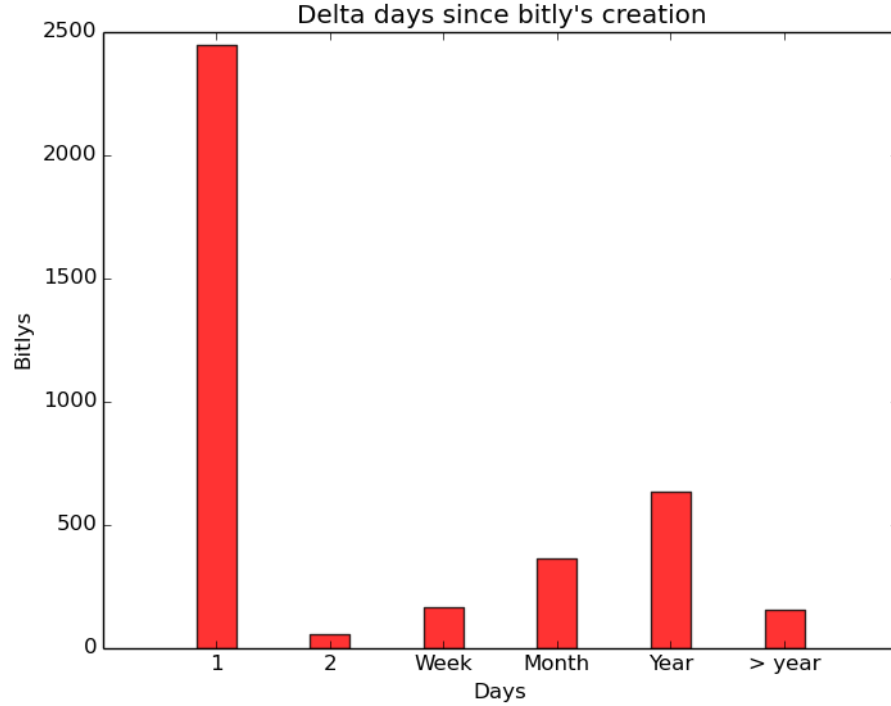
Figure 30. Delta days between creation and tweeting in the collected sample

public timeline to analyze and we utilized the Bitly API to extract the creation date of said bitly. We observed that the creation dates range from one day to several years but the majority of bitlys shared in the present have been created within the last day as shown in Figure 30.

With this knowledge we proceed in our analysis by extracting another dataset of 1,000 random unique tweet-bitly pairs from the Twitter timeline where the bitlys in the tweets have been created a couple of hours from the beginning of the experiment to ensure freshness of the resource referred to by each bitly. This freshness measure is an implicit indicator of the novelty of the resource, as the purpose of this experiment is to capture the lifetime of a resource from its creation and posting to social media and the witnessed changes on that resource.

We conduct an initial analysis on the URIs in the dataset to have a better understanding of the problem. For each URI we recorded the "depth" of that resource (indicated by how many "/" are in the URI), the domain name, and the corresponding category of this domain extracted from the web analytics website Alexa.com. Figure 31 shows the distribution of the depths of the resources in the

dataset. Table 11 shows the top occurring domain names in the dataset. Table 12 shows the domains' categories.
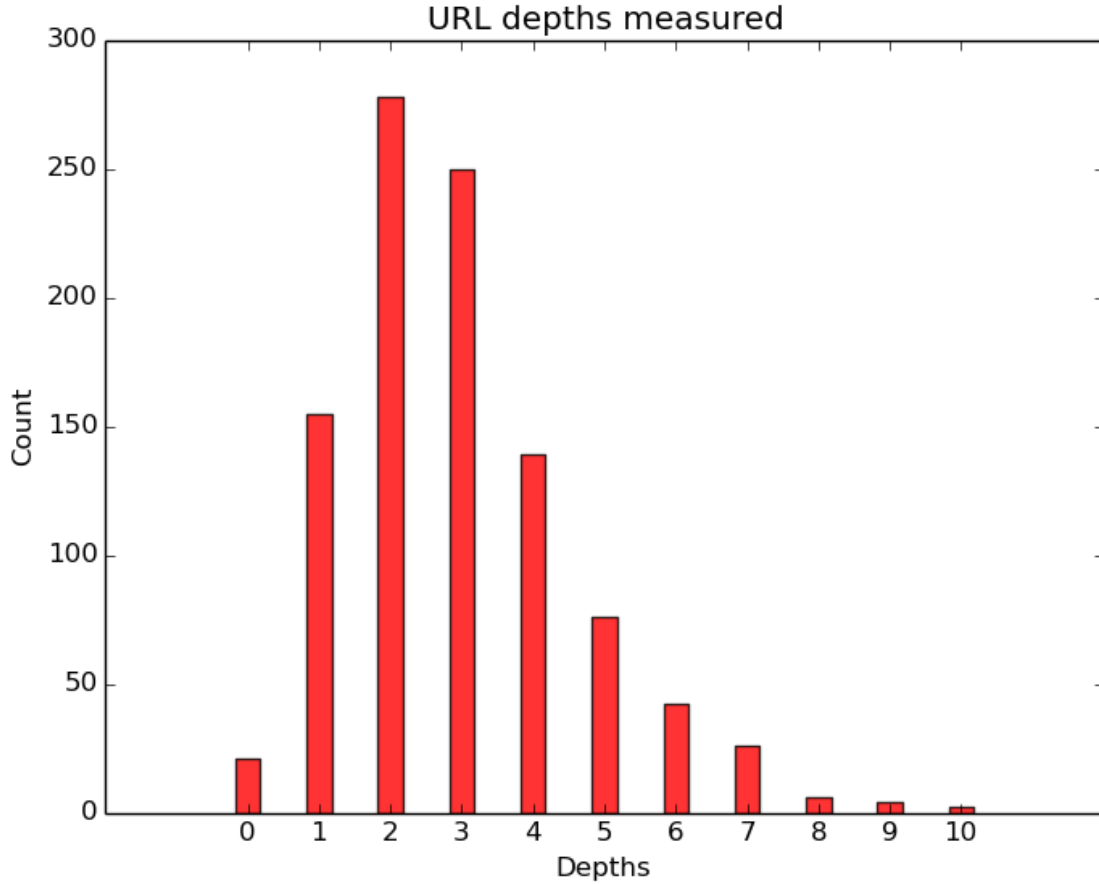
Figure 31. URI depths as they appear in the dataset, (n=1,000)

With this understanding of the collected dataset of the 1,000 resources we start our periodic collection of information related to each instance. For each of the resources in the collection for this longitudinal study we record all the changing information we can capture as follows:

- **From the content aspect**, we download each resource periodically every 45 minutes to capture every change occurring to the content in real time.

- **From social spread aspect**, each hour we record all the tweets posted incorporating a link to each of the resources which highlights the sharing and spread on Twitter.

Rank	Domain	# of appearances in dataset
1	imdb.com	16
2	yahoo.com	7
3	nba.com	6
4	indiatimes.com	4
5	wikipedia.org	4
6	mozilla.org	4
7	google.com	4
8	nih.gov	3
9	about.com	3
10	cnn.com	3
11	nytimes.com	3

Table 11. URI counts based on common domain names in the dataset

Rank	Domain Category	# of appearances in dataset
1	World	51
2	Science/Technology	30
3	Games	22
4	Business	18
5	Shopping/Classifieds	17
6	Society/Paranormal/Organizations	16
7	Arts/Movies/Databases	16
8	Business/Resources/Conferences	16
9	Computers/Programming/	16
10	Sports/Soccer	14
11	Reference/Maps	14
12	Society/Islam	13
13	Computers/Internet	12
14	Reference/Libraries/Research	12
15	News	12

Table 12. Top categories of the domains in the dataset. Categories extracted from Alexa.com

- **Also from the social aspect**, we record the Facebook shares, likes, posts, and clicks once a day.

- **From the activity aspect**, we record the click logs of the bitly using the Bitly API to highlight the activity patterns of this resource and extrapolate the rate of its spread, when it was clicked, read, and shared.

Several static properties are collected as well, like the depth of the resource, length of the unshortened original URI, estimated age of the target resource (using Carbon Date which will be described in Section 5.3), shortening date, and number of mementos in the archives.

To perform this over a long period of time reliably and consistently we decided to utilize Amazon's Web Services (AWS) to deploy our data collection code. We utilized initially a large M3 EC2 Instance with two High Frequency Intel Xeon E5-2670 v2 (Ivy Bridge) processors, 32 GB SSD-based instance storage for fast I/O performance. For data storage, we utilize AWS's S3 buckets that are flexible in size and accessible through the cloud. Initial estimates suggest an average size of one megabyte per snapshot (HTML snapshot, rendered PNG image of the page, topsy tweets collected so far). We capture a snapshot of the URI in batches and each batch takes about 45 minutes to be completed and restarted. This means we have a snapshot of the resource every 45 minutes on average, and are able to collect (24x60)/45 = 32 snapshots per day. Following Equation 7, we estimate an average of one TB of data collected monthly from the 1,000 URIs dataset assuming none disappear. S3 elastic storage can easily accommodate this data size. We present this estimate as it will give us an insight of the cost in processing time and storage rental for the extended period of time (aiming for 6-12 months).

$$Snapshots\ size\ per\ month = 1000\ resources \times 1\ MB \times 32\ snapshots \times 30\ days$$
$$\approx 0.96\ terabyte$$

(7)

We run the code utilizing Amazon's Simple Workflow Service (SWF), which spawns ten concurrent activities, each running our code for data acquisition. Using the workflow we collect a snapshot of each resource in our initial 1,000 URIs dataset every 44.57 minutes on average. The reason we utilize AWS in our experiment is because it is scalable, cheap, easy to deploy, has auto monitoring and logging

(a) After an hour

(b) After six hours

(c) After half a day

(d) After a day

(e) After two days

(f) After a week

(g) After a month

(h) After 52 days
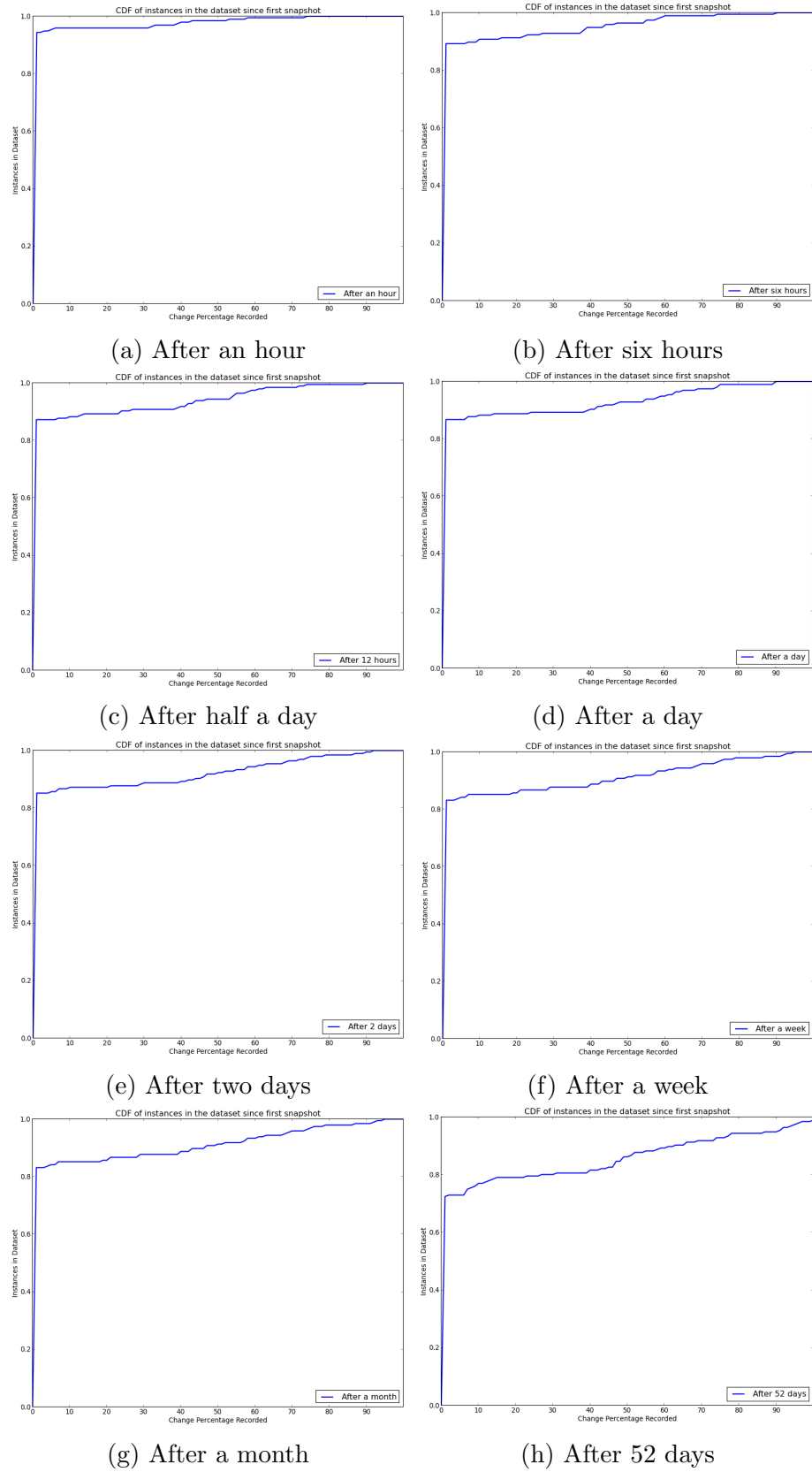
Figure 32. CDFs of the dataset for each time interval, (n=1,000)

utilities, can be programmed to perform auto notifications and handle workflow failures. In this experiment, our group's alumnus, Moustafa Aly, who is currently working at Amazon, provided the experiment and workflow design, and he is helping us maintain the experiment for the upcoming months.

We ran the pilot experiment on AWS on the same exact dataset for 52 days. Unfortunately due to intermittent workflow failure, the snapshots are collected on-and-off during this period yielding 338 snapshots along with their corresponding timestamps. Since we have the downloaded HTML content, we removed the boilerplate and extracted the main textual content. Using a rooted change calculation, we measured the cosine similarity in textual content between the original and the snapshots. Also to calculate the change we subtract 1 - similarity. We record our normalized observations and calculate a cumulative distribution function (CDF) for each time delta since t_{tweet}: one hour, six hours, 12 hours, one day, three days, one week, one month, and all 52 days of the pilot experiment. Figure 32 displays the CDFs for the dataset for each time interval and Figure 33 superimpose them on top of each other for comparison.

From the CDFs we proved our intuition that some shared resources change rapidly within the first hours/days of first sharing on the web. After just one hour, $\sim 4\%$ of the resources have changed by 30%. After six hours, the percentage doubled to be $\sim 8\%$ changed by 40%. After a day the change rate slowed to be $\sim 12\%$ of the resources changed by 40%, while it almost stabilizes after one week at $\sim 17\%$ of the resources to be changed by 40%. This is a rather conservative/optimistic indication of change as we only account for change in the textual content of the resource after removing boiler plate. In reality, this percentage would be higher if we account to the resources that change drastically in the visually-rendered content with only minor HTML changes. A well-known example of small changes in the HTML with semantically significant changes in the reader's perception is that of Google's "doodle's" (some of which are shown in Figure 34). A small change in the HTML at `google.com` to switch the doodle will result in the user experiencing a different commemoration, celebration, etc.

Social media is thought to be disposable and instantaneous. This proved to be far from right as several researchers utilize tweet collections related to events and such as we highlighted in Section 3.1.2. Furthermore, even though the majority of shared links were created within a day of tweeting, it is evident that users also share
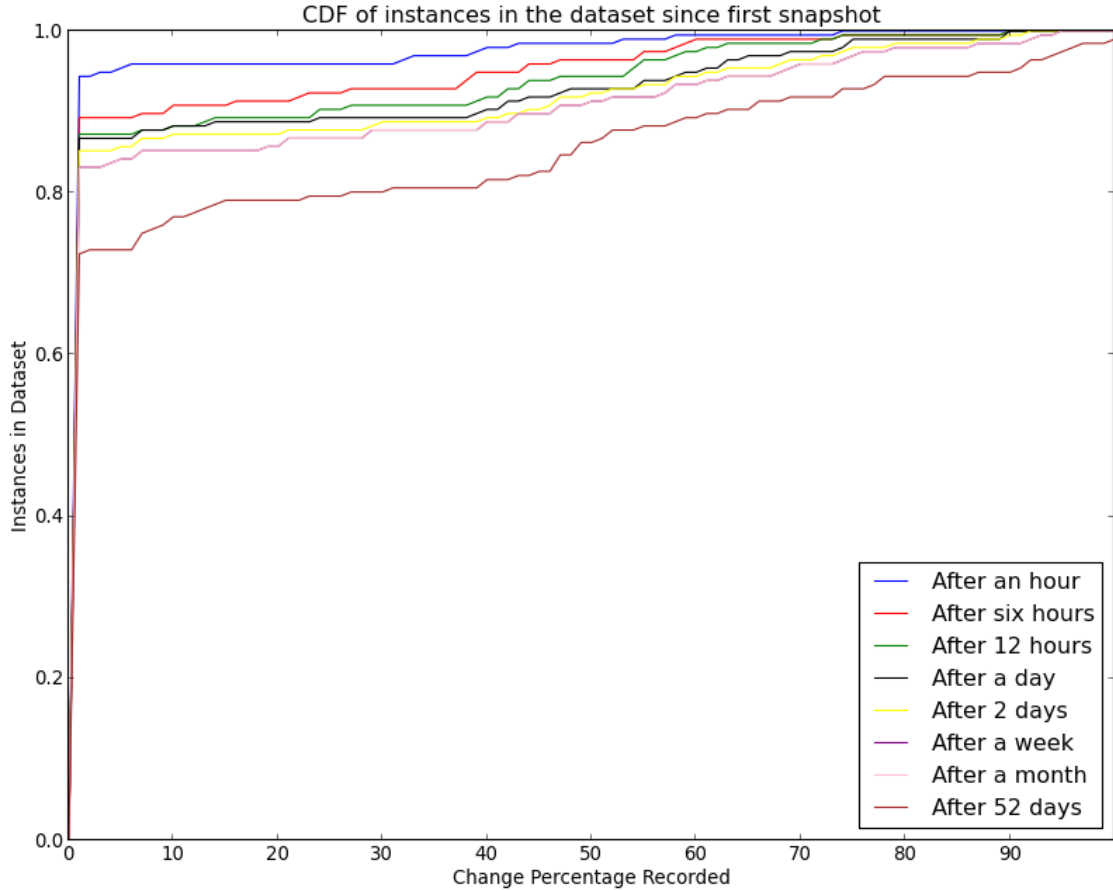
Figure 33. CDF of the dataset with superimposed time intervals, (n=1,000)

much older content. This shows that users at one point or another have shared incorrect content without knowing.

5.2 ESTIMATING WEB ARCHIVING COVERAGE

In order to estimate the ability of the web archives to provide versions of the resource shared in social networks, we had to estimate the archival coverage. To address this, in 2010 we sampled 4000 URIs and measured their coverage in the public web archives and the density of this coverage if it exists [209]. We sampled URIs from DMOZ, Delicious, Bitly, and search engine indices. The search engine indices were randomly sampled using the technique of Bar-Yossef and Gurevich which attempts to remove the search engine bias towards "popular" resources [210]. The results indicate that the source of the URI plays an important role in how much it is archived as shown in Table 13. The experiment is described in further detail

(a) Mahmoud Mokhtar's 121st Birthday. (Egypt 2012)

(b) Mother's Day 2014 (UK)

(c) Amelia Earhart's 115th Birthday (2012)

Figure 34.   Three examples from Google's Doodle page, low HTML change but drastic visual change

in AlSum's doctoral dissertation, as well as the results of revisiting the experiment again in 2013 [211].

| Source | 2010 | | 2013 |
	Including SE Cache	Excluding SE Cache	General
DMOZ	90%	79%	90%
Bitly	97%	68%	95%
Delicious	35%	16%	52%
Search Engine	88%	19%	33%

Table 13. Percentage archived from the web according to source in 2010 and 2013

As much as it is an optimistic notion to have from 33%-90% of the web to be archived this analysis does not address two important and crucial aspects: these percentages are of the indexed web, what about the social web? How well is this archived?

To address the first question AlNoamany et al. analyzed several social media collections and she found that only 12.6% of the resources shared in social media were archived [212]. This shows that the 2011 dataset only covers the indexed web but not necessarily the social web, which is characterized by being much more dynamic in nature, thus showing that the 2011 dataset is not representative of what is shared on social media. Furthermore, since this dataset has been utilized in research since 2011, the URIs with inherently became more exposed and indexed which ease their discovery and become more likely to be archived. In essence, the 2011 data set is a best case, optimistic scenario.

To address the second question, the 2011 study merely checks existence in TimeMaps, not whether or not the page had been archived "well". Brunelle et al. conducted a study in 2014 to gauge how well the resources have been archived and how to calculate damage if it existed in the archived versions [213]. They showed that some embedded resources which are found missing from the archived memento are more significant and should be weighted more heavily than others when computing this damage.

5.3 CARBON DATING THE WEB

In the course of this research, we often needed to compute the creation time of a URI. In this section, we describe the work we published in Carbon Dating the Web [214, 215].

In some webpages, specifically news articles, there is a human readable timestamp indicating when this resource was created or first made available to the public. Unfortunately, this creation timestamp is not available in all webpages. Also, for those select few pages, the timestamp format and location varies largely on the site design, language, orientation, along with the time granularity. Some forum posts could deliver solely the month and the year of the post, while some news sites provide the timestamp to the second. For example, Figure 36a shows the timestamp in a CNN.com page having the timezone, date, and time to the minute. While in ahram.org.eg, and as shown in Figure 36b for a similar article, there is no timezone or time, just the date. Time zones could be problematic too: if not clearly stated on the page, the time zone could be that of the webserver, the client, or GMT.
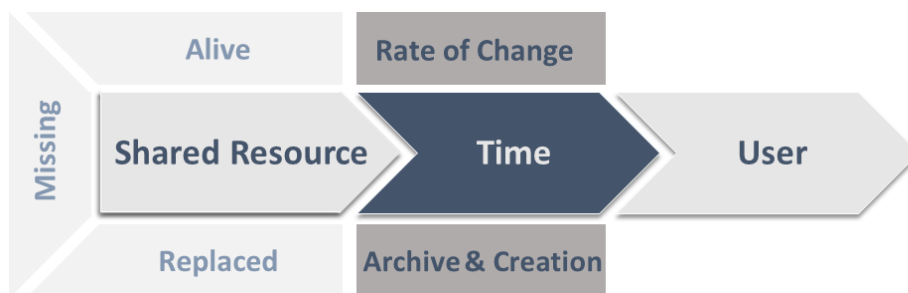
Figure 35. Analyzing the past web: Resource's archived percentage and creation dates

Ideally, each resource should be accompanied by a creation date timestamp. Modern content management systems might keep track of Creation Datetime, but it is not formally defined at the HTTP level, as discussed by Michael Nelson [216]. A second resort would be to ask the hosting web server to return the last-modified HTTP response header. Unfortunately, a large number of servers deliberately return more current last-modified dates to persuade the search engine crawlers to continuously crawl the hosted pages, as shown later in Figure 38b. This renders the dates

(a) Timestamp in a CNN.com article: 3:18 PM ET, Thu March 5, 2015

(b) Timestamp in a Ahram.org.eg article: Thursday, 5 Mar 2015

Figure 36. Timestamps in articles

obtained from the resource highly unreliable.

As discussed in Section 4.1, some of the social media resources we were investigating ceased to exist. We needed to investigate the time line of this resource from creation, to sharing, to deletion. Depending on the hosting server to provide metadata about a missing resource is unachievable in most cases. This places a limitation on services that attempt to parse the resource's textual representation to determine the creation date.

The next step would be to search the public archives for the first existence of the resource. As we show below, using this method solely has significant limitations. Thus there is a need for a tool that can estimate the creation date of any resource investigated without relying on the infrastructure of the hosting web server or the state of the resource itself. Some pages are associated with APIs or tools to extract metadata, but unfortunately these APIs are not standarized and highly specific, and what works on one page would not necessarily work on the other.

Due to the speed of web content creation and the ease of publishing, we make a simplifying assumption. Although in some cases, like in blogs, a page could be created and edited before it is published to the public, we will assume that the creation and publishing of a resource coincide. If the creation date of the resource is unattainable, then the timestamp of its publishing or release could suffice as an estimate of the creation date of the resource. As fire leaves traces of smoke and ashes, web resources leave traces in references, likes, and backlinks. The events associated with creating those shares, links, likes, and interaction with the URI could act as an estimate as well. Referring back to the example of user @Farrah3m in Section 4.3, even if the image or article is not obtainable we can get a timestamp from the tweet itself, and even if the tweet was deleted we can get the tweet's trail from Topsy, as we showed in Figure 24c. If we have access to these events, the timestamp of the first event could act as a sufficient estimate of the resource's creation date. In this experiment, we investigated using those traces on the web to estimate the creation date of the published resource and we proposed an implementation to this tool based on our analysis to be utilized by researchers.

5.3.1 AGE ESTIMATION METHODS

There are three reasons we cannot use just the web archives to estimate the creation date. First, not all pages are archived as discussed earlier in Section 5.1.
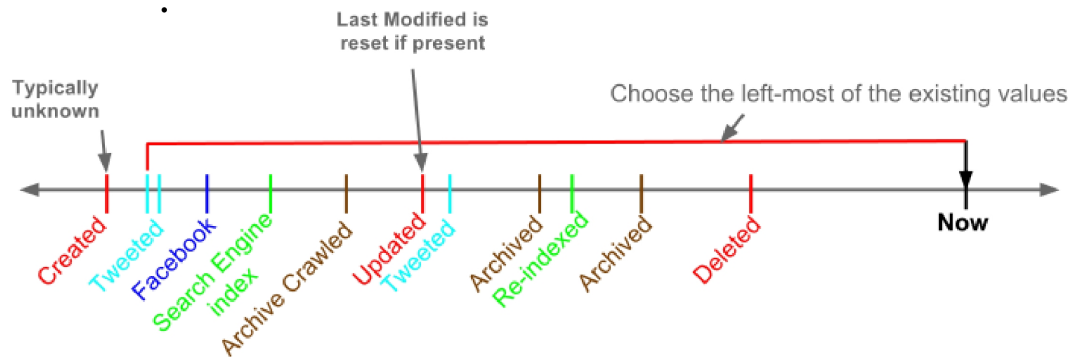
Figure 37. Timeline of typical actions for a shared resource. To estimate the creation date we choose the left-most value

Second, there is often a considerable delay between when the page first appeared and when the page was crawled and archived. Third, web archives often quarantine the release of their holdings until after a certain amount of time has passed (this used to be 6–12 months). Recently, this quarantine period has been eliminated with the "Save a page" feature on archive.org [217].

These reasons limit the use of the web archives in estimating an accurate creation date timestamp for web resources. In the following sections, we investigate several other sources that explore different areas to uncover the traces of the web resources. Utilizing the best of a range of methods, since we cannot rely on one method alone, we build a module that gathers this information and provides a collective estimation of the creation date of the resource. Figure 37 illustrates the methodology of the age estimation process with respect to the timeline of the resource. In this figure, assuming that shortly after the resource's creation it gets tweeted, then Facebook shared, then the search engines add it to their index. Following that, the resource gets archived, changed, or maybe deleted. The tweets, Facebook posts, and other indications of its existence still persist. So we choose the earliest indication of the resource's existence, which would serve as an approximation of the creation date.

Resource and Server Analysis

Prior to investigating any of the web traces, we examine the metadata of the resource itself. We send a HTTP HEAD request to the hosting server and search for the existence of last-modified date response header and parse the timestamp

(a) The article was published on February 12th 2012

```
curl -I http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-
revolution-year.html

HTTP/1.1 200 OK
Content-Type: text/html; charset=UTF-8
Expires: Sat, 02 Mar 2013 04:04:09 GMT
Date: Sat, 02 Mar 2013 04:04:09 GMT
Cache-Control: private, max-age=0
Last-Modified: Wed, 27 Feb 2013 17:27:20 GMT
ETag: "473ba56b-fd4a-4778-b721-3eabdd34154e"
X-Content-Type-Options: nosniff
X-XSS-Protection: 1; mode=block
Content-Length: 0
Server: GSE
```

(b) HTTP response headers displaying last-modified date field

Figure 38. Last modified date example

associated if it exists. We use the `curl` command to request the headers, as shown in Figure 38. We also note the timestamp obtained from the headers can have errors, as demonstrated in a study of the quality of Etags and last-modified datestamps by Clausen [110]. Unfortunately, last-modified response headers are increasingly unavailable because modern content management systems do not provide them.

Backlink Analysis

Typically backlinks are discoverable through search engines. In the next sections we explore the different forms of backlinks and how we can utilize them in our investigation.

Search Engine Backlinks Referring back to the definition of backlinks in Section 2.3, page A has a link on it referring to the intended page B. If Page A is static and never changed this means that it was created at a point in time following the creation of B, which could be by minutes or years. If page A was change-prone and had several versions, the first appearance of the link to page B on A could trigger the same event, indicating that that it happened also at a point in time following the creation of B. If we can search the different versions of A throughout time, we can estimate this backlink timestamp.

To accomplish this, we utilized Google API[1] in extracting the backlinks of the URI. Note that the Google API is known to under-report backlinks, as shown by McCown and Nelson [218]. To explore the multiple versions of each of the backlinks, we utilize the Memento framework in accessing the multiple public archives available [7]. For each backlink we extract its corresponding TimeMaps. We use binary search to discover in the TimeMaps the first appearance of the link to the investigated resource in the backlink pages. Using binary search ensures the speedy performance of this section of the age estimating module. With the backlink having the most archived snapshots (CNN.com > 23,000 mementos), the process took less than 15 iterations accessing the web archives. The earliest of the first appearance timestamps from all the backlinks is selected as the estimated backlink creation date, and this date can act as a good estimation of the creation date of the resource.

Social Media Backlinks Similarly, we follow the definition of a social media backlink as stated in the background chapter, and we argue we can utilize it in identifying creation dates. To elaborate, we examine the following scenario. A resource has been

---

[1]https://developers.google.com/custom-search/v1/overview

Figure 39. Resource published at time $t_{creation} = 2012{:}02{:}11$

created at time $t_{creation}$, as shown in Figure 39 and shortly after a social media post, or a tweet, has been published referring to the resource at time $t_{tweet} = 2012{:}02{:}12$ as shown in Figure 40. This new time, $t_{tweet} = 2012{:}02{:}12T06{:}33{:}00$, could act as a fairly close estimate to the creation date of the post with a tolerable margin of error of minutes in some cases between the original $t_{creation}$ and t_{tweet}.

Figure 40. A tweet posted referencing the resource at time $t_{tweet} = 2012{:}02{:}12T06{:}33{:}00$

Given this scenario, tweets inherently are published with a creation date which makes it easier to extract. The task remaining is to find the tweets that were published with the targeted resource embedded in the text with incorporating all

the shortened versions of the URI as well. Twitter's timeline search facility and its API both provide results of a maximum of nine days from the current day as of 2013 [219]. Accordingly, we utilize another service, Topsy.com, that enables the user to search for a certain URI and get the latest tweets that contained the URI and the influential users sharing it. Topsy's Otter API provides up to 500 of the most recent tweets published embedded a link to the resource and the total number of tweets ever published. Except for highly popular resources, the 500 tweets limit is often sufficient for most resources. The tweets are collected and the corresponding posting timestamps are extracted. The earliest of these timestamps either is or estimates the first time the resource was tweeted. This timestamp in turn signifies the intended $t_{creation}$ mentioned earlier.

Figure 41. BBC.co.uk general public bitly, bit.ly/4Er8c

URI Shortening Backlinks Another form backlinks could take is URI shortnening. Currently, there are hundreds of services that enables the user to create a short URI that redirects to the original longer URI and allows for easier dissemination on the

web. Shortened URIs could be used for the purposes of customizing the URI or for monitoring the resource by logging the amount of times the short URI have been dereferenced or clicked [61]. Some services, like Bitly, can provide the users with a lookup capability for long URIs. When a URI is shortened for the first time by a non logged-in user, it creates an aggregate public short URI that is public to everyone, as shown in Figure 41 (which shows bit.ly/4Er8c as the public shortened URI for BBC.co.uk). When other unauthenticated users attempt to shorten the same URI, it provides the original first aggregated short URI. For every logged-in user, the service provides the possibility to create another personal shortened URI, as shown in Figure 42 (which shows user heinestien's personal bitly, bit.ly/1MbRwwU). For our purposes, we lookup the aggregated short URI indicating the first time the resource's URI has been shortened by this service and from that we query the service once more for the short URI creation timestamp. Bitly was used as the official automatic shortener for period of time by Twitter before Twitter replaced it with their own shortener, t.co, in 2010. Similarly to the previous backlinks method, we mine Bitly for those creation timestamps and use them as an estimate of the creation date of the resource, assuming the author shortens and shares the resource's URI shortly after publishing it.

Archiving Analysis

The most straightforward approach used in the age estimation module is the web archives analysis. We utilize the Memento framework to obtain the TimeMap of the resource, from which we extract the earliest Memento-datetime. Note that Memento-datetime is the time of capture at the web archive and is not equivalent to last-modified or creation date [216]. In some cases, the original headers in some mementos include the original last-modified dates, but all of them have the Memento-datetime fields. We extract each of those fields, parse the corresponding dates, and pick the earliest. An extra filter was added to avoid dates prior to 1995, before the Internet Archive began archiving, or datestamps greater than the current timestamp.

Search Engine Indexing Analysis

The final approach is to investigate the search engines and extract the last crawled date. Except for highly active and dynamic web pages, resources are crawled
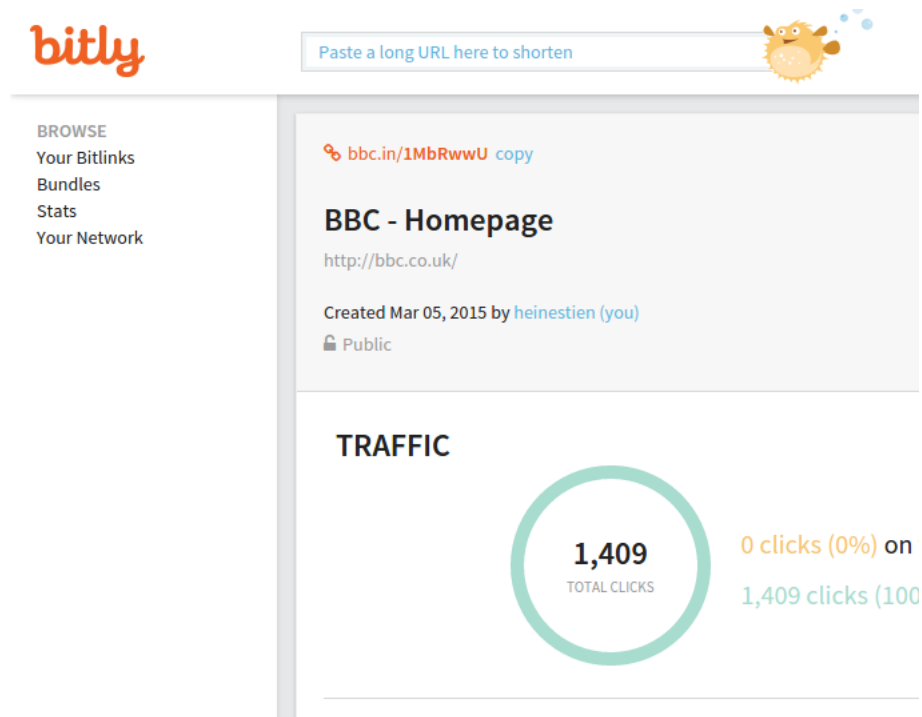
Figure 42. BBC.co.uk personal bitly, bit.ly/1MbRwwU created after logging in

once and marked as such to prevent unnecessary re-crawling [220]. News sites article pages, blogs, and videos are the most common examples. The idea is to use the search engines' APIs to extract this last crawled date and utilize it as an estimate of the creation date. This approach is effective due to the relatively short period of time between publishing a resource and its discovery by search engine crawlers. We use Google's search API and modify it to show the results from the last 15 years accompanied by the first crawl date. Unfortunately, this approach does not give time granularity (HH:MM:SS), just dates (YYYY:MM:DD).

5.3.2 ESTIMATED AGE VERIFICATION

To validate an implementation of the methods described above, we created a gold standard dataset from different sources from which we can extract the real publishing timestamps. This could be done by parsing feeds, parsing web templates, and other methods. In the next sections we illustrate each of the sources utilized and explain

the extraction process.

Gold Standard Data Collection

Two factors were crucial in the data collection process: the quality of the timestamps extracted, and the variety of the sources to reduce any bias in the experiment. Thus, we divide data into four categories: news sites, social media sites, Alexa.com's top domains, and manual extraction. Table 14 summarizes the four categories.

	Data Sources	Resources Collected	Sampled Resources	Timestamp Allocation Method
News Sites	news.Google.com	29,154	100	XML sitemap
	BBC.co.uk	3,703	100	Page Scraping
	CNN.com	18,519	100	Page Scraping
	news.Yahoo.com	34,588	100	XML sitemap
	theHollywoodGossip.com	6,859	100	Page Scraping
Social Sites	Pinterest.com	55,463	100	RSS feed
	Tumblr.com	52,513	100	RSS feed
	Youtube.com	78,000	100	Search API
	WordPress.com	2,405,901	100	Atom feed
	Blogger.com	32,417	100	Atom feed
	Alexa.com Top Domains	167	100	Page Scraping & Who.is service
	Manual Extraction	100	100	Manual inspection
	Total:	2,717,384	**1,200**	

Table 14. The resources extracted with timestamps from the web forming the gold standard dataset

News Sites Each article is associated with a timestamp in a known template that can be parsed and extracted. The articles are also usually easily accessible through RSS and Atom feeds or XML-sitemaps. For each of the news sites under investigation, we extracted many resources then randomly downsized the sample.

Social Media and Blogs To increase the variety of the gold standard dataset, we investigated five different social media sources. These selected sources are highly popular and it is possible to extract accurate publishing timestamps. As those sources are tightly coupled with the degree of popularity and to avoid the bias resulting from this popularity we randomly extract as many resources as possible from the indexes, feeds, and sitemaps and do not rely solely on the most famous blogs or most shared tumblr posts. Furthermore, we randomly and uniformly sample

each collection to reduce its size for our experiment.

Long Standing Domains So as not to limit our gold standard dataset to low level articles, blogs, or posts only, we incorporated long-standing top-level domains. To extract a list of those domains, we mined Alexa.com for the list of the top 500 sites[2]. This list of sites was in turn investigated for the DNS registry dates using one of the DNS lookup tools available online, as shown in Figure 43. For these domain names, we assume the existence of a site (with www. prepended) that corresponds with the domain name. A final set of 100 was randomly selected from the resolved sites and added to the gold standard dataset.

Figure 43. Pinterest.com (Alexa global rank = 37), registered on 26th November 2009, released March 2010

Manual Random Extraction Finally, we randomly select a set of 100 URIs that we can visually identify the timestamp somewhere on the page itself. These URIs were selected empirically using random walks on the web. The ten URIs analyzed [221] are included within these 100 URIs along with their corresponding creation timestamps. The corresponding true value of the creation timestamp for each of the ten URIs is the one provided in their analysis.

---

[2]http://www.alexa.com/topsites

Experimental Analysis

The collected dataset of 1,200 pairs of URIs and manually verified creation dates was tested against an implementation of the carbon dating methods. Since the data came from different sources, the granularity varied in some cases, as well as the corresponding time zones. To be consistent, each real creation date timestamp t_{real} was transformed from the corresponding extracted timestamp to Coordinated Universal Time (UTC) and has been truncated to ignore the time portion and keep just the date. Each data point has a real creation date in the ISO 8601 date format without the time portion (e.g., YYYY:MM:DD). Similarly, the extracted estimations were processed in the same manner and recorded.

For each method, we recorded the estimated timestamp t_{method} and the temporal delta Δt_{method} between t_{method} and t_{real}, as shown in Equation 8. Collectively, we calculate the best estimated timestamp $t_{estimated}$ as in Equation 9, the closest delta between all the methods Δt_{least} and the real timestamp t_{real}, as shown in Equation 10, and the method that provided this best estimate.

$$\Delta t_{method} = |t_{real} - t_{method}| \tag{8}$$

$$t_{estimated} = min(t_{method}) \tag{9}$$

$$\Delta t_{least} = |t_{real} - t_{estimated}| \tag{10}$$

Table 15 shows the outcomes of the experiment. The numbers indicate how many times a resource provided the closest timestamp to the real one. It also shows that for 290 resources (24.90%), the module failed to provide a creation date estimate.

5.3.3 CREATION DATE EVALUATION

As our age estimation module relies on other services to function (e.g., Bitly, Topsy, Google, Web archives), the next step is to measure the effect of each of the six different age estimation methods and to gauge the consequences in failure to obtain results from each. For each resource, we got the resulting best estimation and calculated the distance between it and the real creation date. We set the granularity of the delta to be in days to match the real dates in the gold standard dataset. To elaborate, if the resource was created on a certain date and the estimation module returned a timestamp on the same day we declare a match and in this case Δt_{least}

= 0.

Age Estimation Method	Using Best Estimate		Contribution	
	Number Of Resources Found	Percentage Of Resources Found	Resources Contributed	Percentage Contributed
Bitly	96	10.55%	554	46.21%
Google	370	40.66%	709	59.13%
Topsy	236	25.93%	632	52.71%
Archives	152	16.70%	578	48.21%
Backlinks	3	0.33%	180	15.01%
Last-Modified	53	5.82%	134	11.18%
Total Estimate	**910**	**75.90%**	1199	100%

Table 15. Results of testing the gold standard dataset against the six age estimation methods (n=1200)

Method of Estimation	Area Under Curve (AUC)	Percentage lost in AUC
Bitly	758.73	0.51%
Google	742.52	2.64%
Topsy	720.61	5.51%
Archives	741.23	2.81%
Backlinks	762.64	0%
Last-Modified	725.59	4.46%
Total Estimate	762.64	0%

Table 16. Area under the curve for the six age estimation methods

To measure the accuracy of estimation, 393 resources out of 1200 (32.78%) returned $\Delta t_{least} = 0$ indicating a perfect estimation. For all the resources, we sorted the resulting deltas and plot them. We calculated the area under the curve using the composite trapezoidal rule and the composite Simpson's rule with x-axis spacing of 0.0001 units. We took the average of both approximations to represent the area under the curve (AUC). Semantically, this area signifies the error resulting from the estimation process. Ideally, if the module produced a perfect match to the real dates, AUC = 0. Table 16 shows that the AUC using the best lowest estimate of all the six methods is 762.64. Disabling each method one by one and measuring the AUC indicates the resultant error corresponding to the absence of the disabled

method accordingly. The table shows that using or disabling the use of backlinks barely affected the results. Disabling the Bitly services or the Google search index query affected the results slightly (0.51% and 2.64%, respectively), while disabling any of the public archives query, or the social backlinks in Topsy and the extraction of the last-modified date greatly affects the results.

We utilized polynomial fitting functions to fit the values corresponding to the age estimations corresponding to each URI. Figure 44 shows the polynomial curve of the second degree used in fitting the real creation times stamps of the gold standard dataset. Figures 45, 46, 47, 48, and 49 show the fitted curve resulting from removing each of the methods one by one. Each of the curves signifies an estimate of the best the other methods could provide. The further the estimated curve is from the real one, the less accurate this estimation would be.

5.3.4 APPLICATION: CARBON DATE API

After validating the accuracy of the developed module, the next step was to provide age estimation as a public web service. To fulfill this goal, we created "Carbon Date", a web based age estimation API. To use the API, simply concatenate
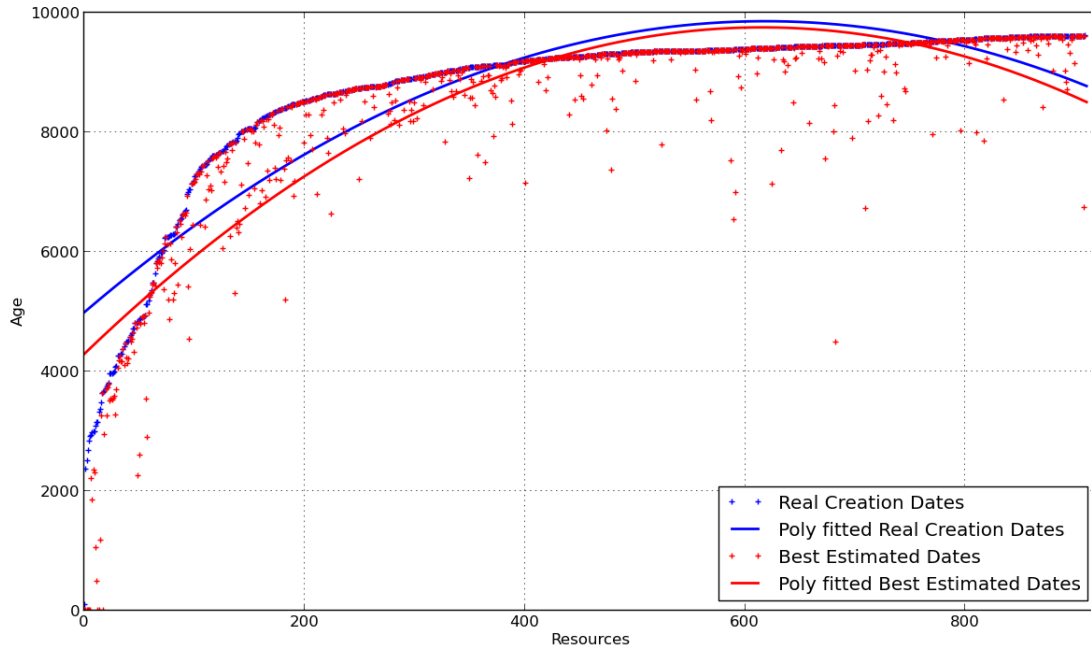
Figure 44. The polynomial fitted curve corresponding to the real creation dates against the estimated creation dates from the module AUC = 762.64
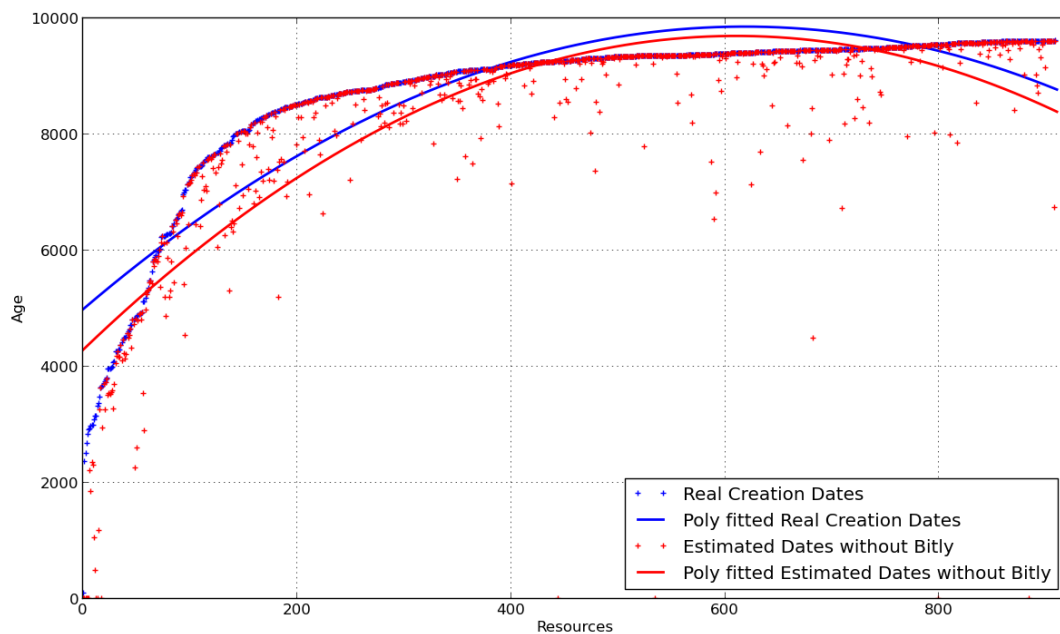
Figure 45. The polynomial fitted curves corresponding to the absence of Bitly, AUC = 758.73
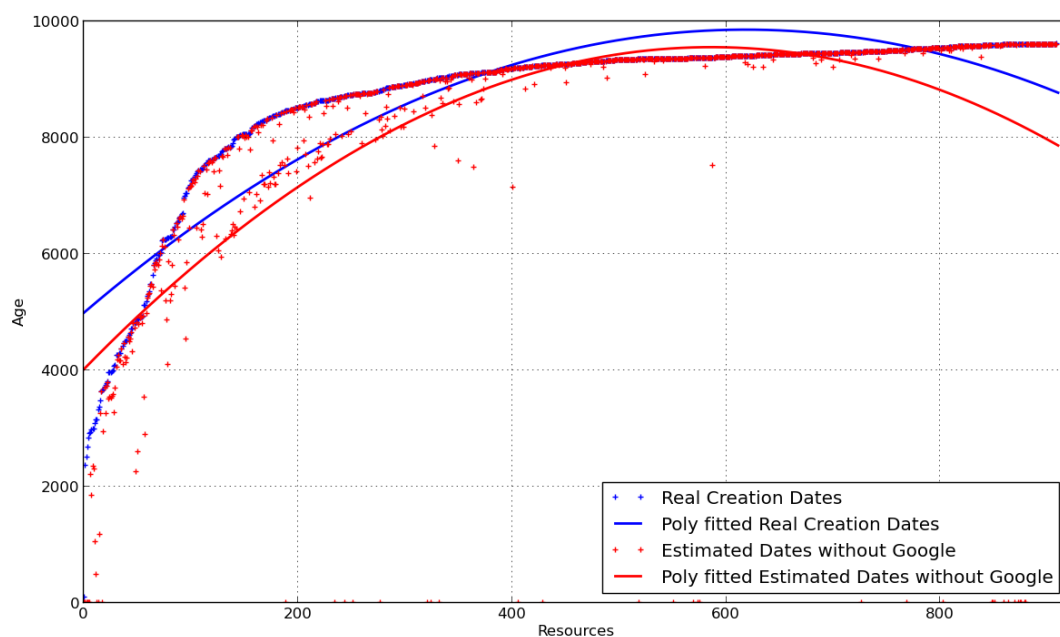
Figure 46. The polynomial fitted curves corresponding to the absence of Google, AUC = 742.52
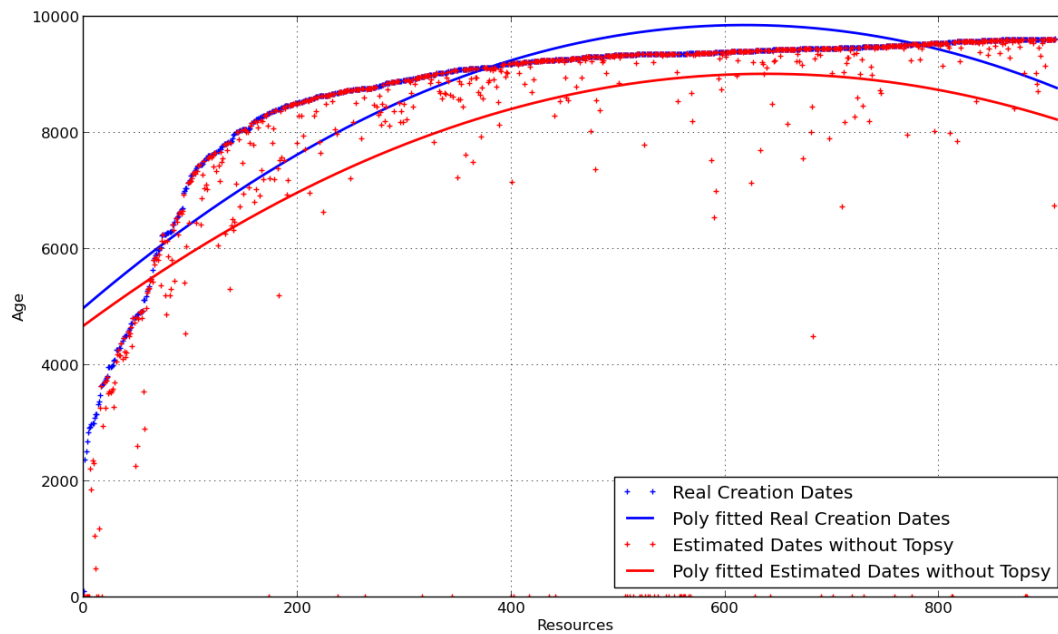
Figure 47. The polynomial fitted curves corresponding to the absence of Topsy, AUC = 720.61
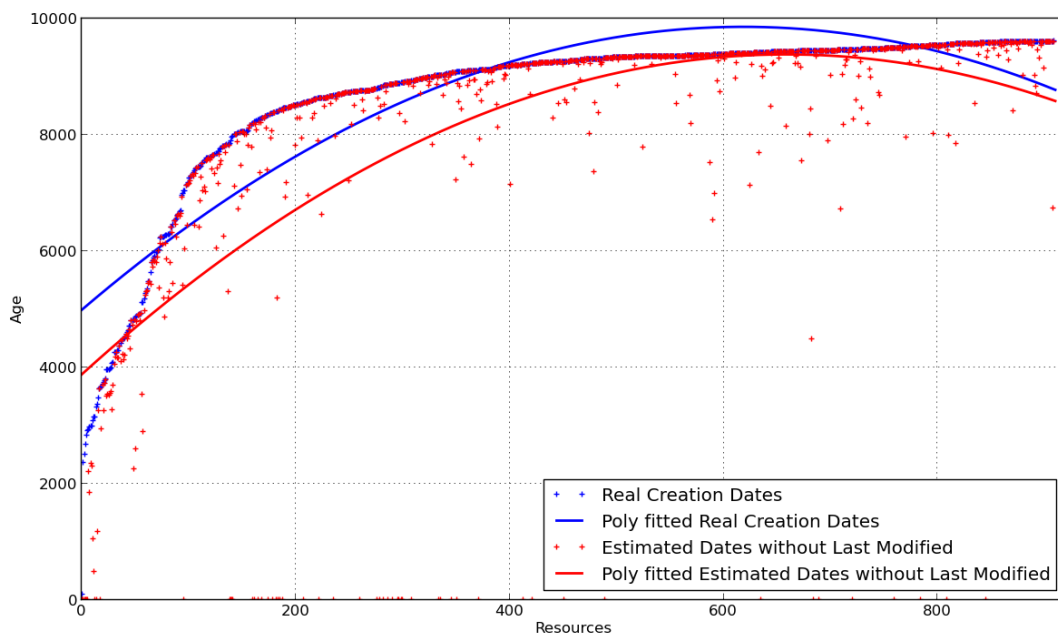
Figure 48. The polynomial fitted curves corresponding to the absence of the Last-Modified, AUC = 725.59
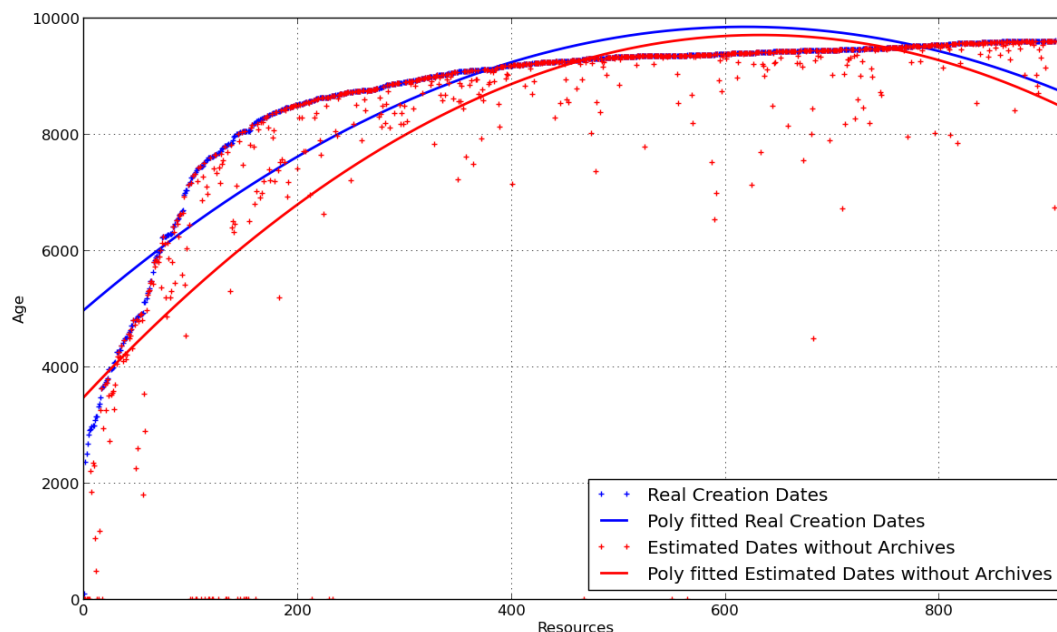
Figure 49. The polynomial fitted curves corresponding to the absence of the Archives, AUC = 741.23

the URI of the desired resource to the following path: `http://cd.cs.odu.edu/cd?url=`. The resulting JSON object would be similar to the one illustrated in Figure 50.

In 2014, Alexander Nwala has developed a second version of Carbon Date and released it to the public [215]. In Carbon Date V2.0, Nwala has addressed the short-comings of the prior version in terms of server caching, multi-threading, optimizing backlinks calculations, and increased the overall efficiency. He also released an in-stallation for a local version that users can set up on their machines. Figure 51 shows Carbon Date's web interface.

5.3.5 SUMMARY

To conclude, previous research investigated the use of public archives as a point of reference to when the content of a certain page appeared. In this experiment, we investigated several other possibilities in estimating the accurate age of a re-source, including social backlinks (social posts and shortened URIs), search engine backlinks, search engine last crawl date, the resource last-modifed date, the first appearance of the link to the resource in its backlinks sites, and the archival first crawl timestamp. We also incorporated the minimum of the original last-modified

```
{
  "self": "http://cd.cs.odu.edu/cd?url=http://www.cnn.com",
  "URI": "http://www.cnn.com",
  "Estimated Creation Date": "1998-12-06T04:02:33",
  "Last Modified": "",
  "Bitly.com": "2008-06-08T12:00:00",
  "Topsy.com": "2015-01-25T23:31:42",
  "Backlinks": "2003-03-12T05:35:44",
  "Google.com": "2005-01-11T00:00:00",
  "Archives": [
    [
      "Earliest",
      "1998-12-06T04:02:33"
    ],
    [
      "By_Archive",
      {
        "http://archive.today/20000815052826/http://www.cnn
        .com/": "2000-08-15T05:28:26",
        "http://arquivo.pt/wayback/wayback/20000815052826/
        http://www.cnn.com/": "2000-08-15T05:28:26",
        "http://wayback.vefsafn.is/wayback/20011106102722/
        http://www.cnn.com/": "1998-12-06T04:02:33",
        "http://web.archive.org/web/20131218180509/http://
        www.cnn.com/": "2013-12-18T18:05:09"
      }
    ]
  ]
}
```

Figure 50. JSON Object resulting from the Carbon Date API. No vote for the "last-modified" key indicates that the HTTP response header did not exist

response header, and the Memento-Datetime HTTP response header. All of these methods combined, where we select the oldest resulting timestamp, proved to provide an accurate estimation to the creation date upon evaluating it against a gold standard dataset of 1200 web pages of known publishing/posting dates. We succeeded in obtaining an estimated creation date to 910 resources out of the 1200 in the dataset (75.90%). Of the closest estimated dates, 40% were obtained from Google. Topsy came in second with 26%, followed by the public archives, Bitly,
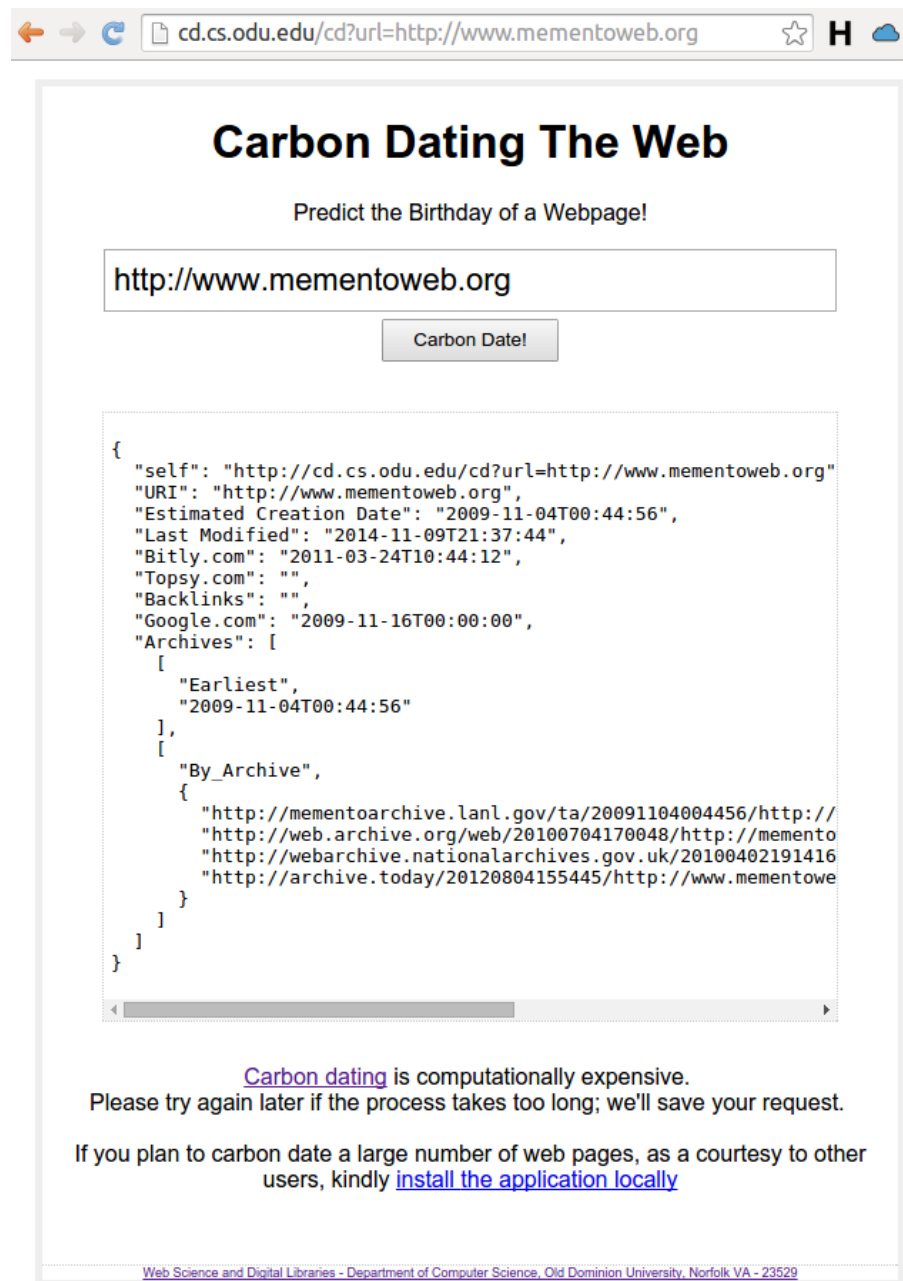
Figure 51. Carbon Date's web interface

and Last-Modified header with 17%, 11%, and 6% respectively. Using the backlinks yielded only three closest creation dates proving its insignificance. We also simulate the failure of each of the six services one at a time and calculated the resulting loss in accuracy. We show that the social media existence (Topsy), the archival existence (Archives), and the last modified date if it exists, are the strongest contributers to

the age estimation module respectively.

CHAPTER 6

USER'S TEMPORAL INTENTION

"Verily, deeds are only with intentions. Verily, every

person will get rewarded only for what they intended."

— Prophet Muhammad, PBUH, *Sahih Bukhari 1*

With a better understanding of content change, persistence, age, and archivability acquired from the previous experiments described in Chapters 4 and 5, we proceed in analyzing the third and final component of our analysis, the user (Figure 52). We commence by defining the meaning of users' intention with respect to time by closely breaking down its proposed components and amass human subjects' interpretation of intention. In this chapter we describe our published work in performing several Mechanical Turk experiments and highlighting the best possible ways to understand and detect intention [222].
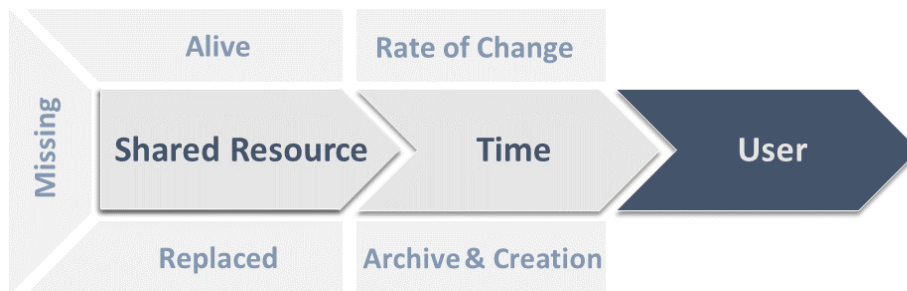
Figure 52. Third analysis component: The user

To have a better understanding of a user's temporal intention, we performed several experiments using Amazon's Mechanical Turk. Subsequently, we discovered that classifying temporal intention is difficult for Mechanical Turk workers. A possible explanation for this observation is that most users are stuck in the web's prepetual now and do not posses an understanding of the concept of time on the

web. Furthermore, intention is highly subjective and hard to describe. This, in turn, has influenced us to seek a simplification of the problem of intention to the more familiar problem of relevancy.

6.1 PRELIMINARY STUDY: HOW NOT TO MEASURE TEMPORAL INTENTION

Initially, in classifying intention, our first set of experiments involved sampling 1000 tweets from the SNAP Twitter data set. The first step was to prove that Mechanical Turk could be used in representing manually assigned classes of intention made by experts in the field. The classes targeted were as follows: did the author of the tweet intend the "Current State" of the resource for the reader at any time or the "Past State" of the resource at the time of the tweet? Or is there not enough information?

To achieve this, we established the ground truth intention for 100 tweets from the set of 1000 tweets forming the gold standard dataset. The intention was determined by polling via email the members of our Web Science and Digital Libraries (WSDL) research group and asking them to classify the intention of the author of a tweet as either the current version (t_{click}), the archived version (past) (t_{tweet}), or unknown by looking at the tweet. The reliability of agreement within our group of 12, all of whom are well aware of the concept of time on the web, web archiving, and the depth of our research question, was surprisingly low (Fleiss' $\kappa = 0.14$). Nonetheless, we ran the same experiment on Mechanical Turk and asked the turkers to choose which version the author intended for the readers to see, and showed them a side by side comparison of the two states of the resource at t_{tweet} and t_{click} as shown in Figure 53. We collected five evaluations for each of the 100 tweets from the gold standard dataset. The inter-rater agreement between the Mechanical Turk workers was even lower (Fleiss' $\kappa = 0.07$).

$$Vote_{MT}(tweet) = \begin{cases} \text{Current,} & \text{if } \dfrac{\Sigma Vote_{current}}{N_{turkers}} > a \\ \text{Past,} & \text{otherwise} \end{cases} \tag{11}$$

The threshold a in Equation 11 defines the agreement vote cut off. In this case, $a = 0.5$ as we applied a simple majority vote in deciding the collective vote of the Mechanical Turk workers (i.e., whichever classification received three out of five votes), and similarly within the 12 WS-DL members. Treating each group as a single

Figure 53. The first Mechanical Turk experiment for intention classification

entity, the aggregated votes from each of the two datasets were used to calculate the inter rater agreement resulting in Cohen's $\kappa = 0.04$, indicating slight agreement. This slight agreement was yet not sufficient to proceed with our study. Examining

the selection from the SNAP data set, we decided that too many of the tweets had vague contexts and were hard to classify.

Given the unclear contexts that were present in the first sample set, we then tried a different dataset from which to sample. We used the tweets from the six historical events described in Section 4.1.1. For 100 tweets, we built a web page with an image snapshot of the current version of the page and a version of the page closest to t_{tweet} that could be found in a public web archive. We held a face to face meeting with our WSDL research group to determine the ground truth: for each tweet we went around the table and argued for whichever version we thought matched the author's temporal intent. We knew this data set would be biased toward t_{tweet} because most of the tweets described historic or cultural events from 2009-2011. After deliberation, we arrived at: 82% past, 9% current, and 9% undecided as our gold standard for this data set. When we submitted the jobs to Mechanical Turk, we defined levels of three, five, seven, and nine evaluations for each tweet. In the case where we had nine evaluations for each tweet, the Mechanical Turk workers would match our gold standard 58% of the time if we allowed 5-4 splits. If we were more discerning and counted agreement only in cases where workers agreed 6-3 or better, then the agreement with Mechanical Turk workers fell to 31% (and similarly for rating levels three, five, and seven).

In short, if we required clear agreement on the part of Mechanical Turk workers, then we did much worse than simply flipping a coin – in a data set with a clear bias toward t_{tweet} because of the focus on past events. It was at this point we decided our approach in discerning the author's temporal intent was simply too complicated for Mechanical Turk workers.
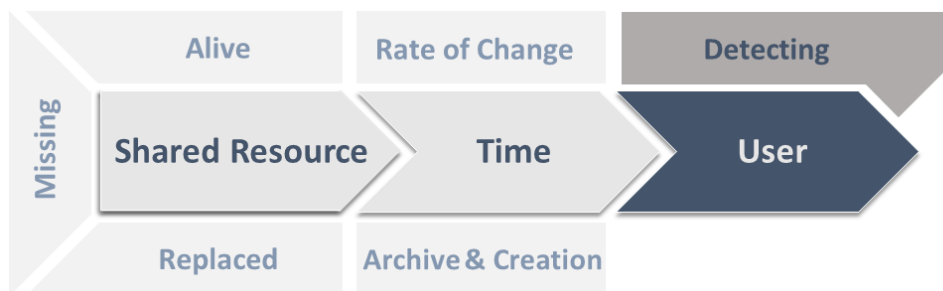
Figure 54. Detecting and understanding user's temporal intention in social media

6.2 TEMPORAL INTENTION RELEVANCY MODEL

To reach our goal in detecting and modeling users' temporal intentions as shown in Figure 54, we need to collect a large dataset, which, as discussed in the previous section, is not a trivial task. The difficulty in acquiring the data resides in generating the ground truth or gold standard for the temporal intention of the user who authored the original social media post. Initially, our intention was to generate a small set of gold standard data (e.g., links classified as representing the user's intention to be either "the resource at t_{tweet}" or "the resource at t_{click}"). After conducting the preliminary study (described in Section 6.1), we decided that the notion of "temporal intention" was too nuanced to be adequately conveyed in the instructions for the workers of Mechanical Turk. From the related works focusing on Mechanical Turk (Section 3.6), it was apparent that turkers excel in categorization and classification tasks, tasks with short descriptions and highly defined smaller tasks at scale. Learning from our previous unsuccessful attempts, we chose to transform the problem of "temporal intention" to a simpler space with two components, one of relevancy between the tweet and the resource as it exists now, and the other of the change amount in textual content. We can calculate the percentage of change using several text-processing techniques and utilize the turkers solely for the relevancy task.

To transform intention into relevancy we examined several of thousands of tweets and their corresponding embedded resources. We first assume we have a resource R which has been tweeted by some author at time t_{tweet}. The state of the resource at t_{tweet} is R_{tweet}. Consequently, another user (the reader) clicked on the resource to read it at a later time t_{click}. The state of the resource at t_{click} is R_{click}. We found that in terms of relevancy and change a tweet-resource pair would typically fall into one of four possible states:

Changed & Relevant: If the resource has changed (i.e. R_{tweet} is not similar to R_{click}) and it is still relevant to the tweet. Figure 55 shows an author tweeting about the latest updates for a newsletter. The linked resource in the tweet continually changes while the tweet is always relevant to it.

Changed & Non-Relevant: If the resource has changed and it is not relevant to the tweet. Figure 56 shows an author tweeting about specific breaking news on

(a) Tweet is still relevant
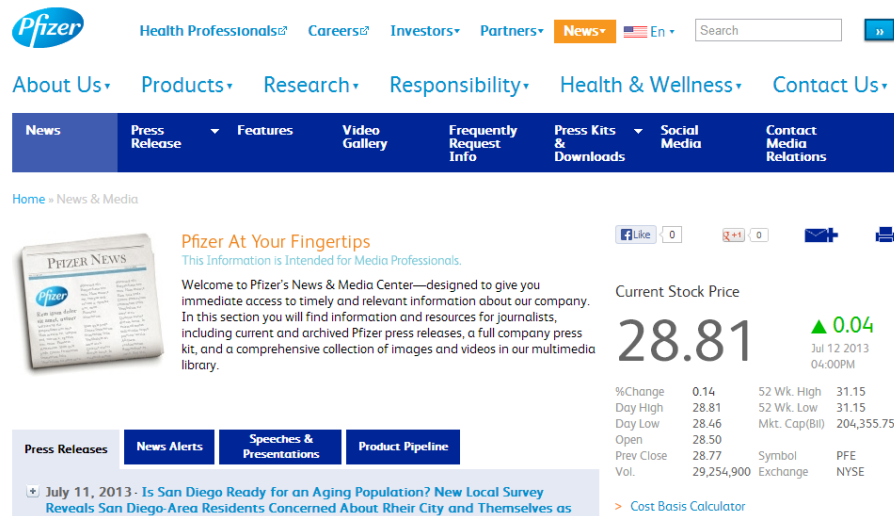
(b) The resource has changed

Figure 55. Resource has changed but is still relevant to the tweet

CNN.com's first page, which by default changes frequently, rendering the resource to be no longer relevant to the tweet.

Not Changed & Relevant: If the resource has not changed and it is still relevant to the tweet. Figure 57 shows an author tweeting about an article which still exists. This is the most common case a user might encounter.

Not Changed & Non-Relevant: If the resource has not changed and it is not relevant to the tweet. Figure 58 shows an author tweeting about a possible spam site. This scenario can occur in spam, mistaken link sharing, or more likely that relevancy relies on out-of-band communication between the original author and the intended readers, for example "Rickrolling"[1].

Given these observations we define our temporal intention model based on change and relevance, the Temporal Intention Relevancy Model (TIRM). We can calculate change based on a multitude of resource similarity algorithms. The key is to assess the relevancy between the tweet and the resource at t_{click}. Figure 59 shows the four cases of TIRM and we can deduce the intended intention as following:

Changed & Relevant: This indicates that the author's temporal intention to be for the *current* version at t_{click}.

Changed & Non-Relevant: The resource has changed and it is not relevant to the tweet; we assume initial relevance and thus the original author must have meant to share the resource in the state as it existed at t_{tweet}, which is R_{tweet} not R_{click}. This indicates that the author's temporal intention to be the *past* version at t_{tweet}.

Not Changed & Relevant: The resource has not changed and it is still relevant to the tweet, then we claim that the intention of the author was to share the resource as it existed at t_{tweet} (R_{tweet}), but it is just a fortunate coincidence that the resource has not changed and is thus still relevant. Since, there is a possibility that the resource could change in the future and become non-relevant, we define the author's intention to be for the *past* version at t_{tweet}.

---

[1]The Internet meme of "Rickrolling" http://en.wikipedia.org/wiki/Rickrolling is a humorous example of purposeful non-relevancy between the context of the link and the link which is to the 1987 pop song by Rick Astley; the point is to "trick" users into expecting one thing and the link delivers the song.

(a) Tweet is no longer relevant

(b) The resource has changed

Figure 56. Resource has changed but is no longer relevant to the tweet

(a) Tweet is still relevant

(b) The resource has not changed

Figure 57. Resource has not changed and is still relevant to the tweet

(a) Tweet is no longer relevant
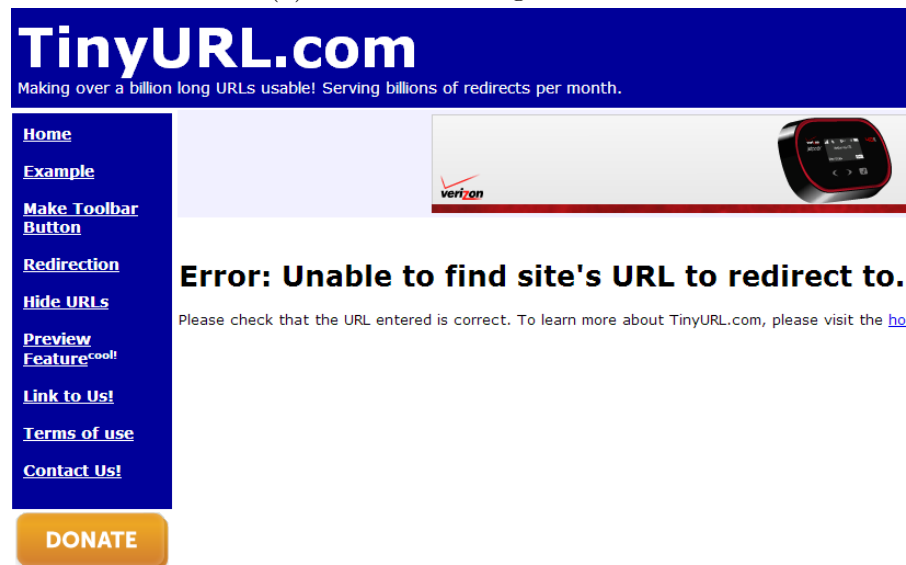
(b) The resource has not changed

Figure 58. Resource has not changed and is not relevant to the tweet

(a) Changed and Relevant

(b) Changed and no longer Relevant

(c) Not Changed and Relevant

(d) Not Changed and not Relevant

Figure 59. Examples of the relevancy mapping of TIRM

Not Changed & Non-Relevant: The resource has not changed and it is not relevant to the tweet, then we can not be sure of the intention and either t_{click} or t_{tweet} will suffice.

		Tweet and resource are:	
		relevant	not relevant
Linked resource has:	changed	t_{click}	t_{tweet}
	not changed	t_{tweet}	either or undefined

Table 17. TIRM: choosing t_{click} or t_{tweet} based on relevancy between the tweet and the resource

Table 17 presents the choice of t_{click} or t_{tweet} based on the assessment of relevance by workers at Mechanical Turk. To resonate with one of the common types of experiments in it, we designed our new experiment as a categorization of relevance problem, which the workers are familiar with. In each Human Intelligence Task (HIT), the worker is presented with the full tweet, its publishing date, and in an embedded window, a snapshot of the page that the tweet links to in its current

state. Instead of asking workers about temporal intention of the original author and possibly confusing it with the temporal intention of them as a reader, we asked a simpler question: "is this page still relevant to this tweet?" There is considerable precedence in the Mechanical Turk community for making relevance judgments as categorization problems are commonly available as HITs and Mechanical Turk by default provide categorization templates in the set of predefined HITs.

6.3 DATASET COLLECTION

After laying the basis of the intention-relevance mapping in TIRM, we must collect a large dataset to be utilized in the modeling and analysis phases. In fact, we collect a proof of concept small dataset first to validate the viability of TIRM to represent temporal intention; then we collect the large dataset to use in training the model.

6.3.1 PROOF OF CONCEPT DATASET

Prior to collecting the training dataset, we need to be confident in the ability of our data collection experiment in representing real-life educated judgement. To achieve this goal, we created a proof of concept dataset by obtaining a small dataset and assigning it to members of our research group, in whom we have confidence of their ability to perform the task accurately, and then assigned the same dataset to workers in Mechanical Turk. We collect both sets of assignments and if the rater agreement was significant, that would indicate the viability of using Mechanical Turk assignments as accurately as we would be utilizing expert opinions. In other words, we can mimic the judgment of the experts and expand in volume. Mechanical Turk HITs are considerably cheaper, easier to acquire, and faster to conclude than the expert assignments.

For the proof of concept dataset, we randomly picked 100 tweets from the SNAP dataset dating back to June 2009 and posted them to be classified as "still relevant" or "no longer relevant". For each HIT we posted the tweet, the date, and a snapshot of the resource at t_{click} (R_{click}). The experiment requested five unique raters with high qualifications (more than 1000 accepted HITs and more than 95% acceptance rate). Each HIT cost two cents and a maximum time span of 20 minutes. The experiment was completed within the first hours from posting and the average completion time per hit was 61 seconds. We examined the data from the workers and dismissed all

Agreement in three or more votes	93%
Agreement in four or more votes	80%
Agreement with all five votes	60%

Table 18. Agreement between the research group and Mechanical Turk workers for 100 tweets

the HITs that took less than 10 seconds, assuming this indicated a hasty decision. We also filtered out workers who exhibited low quality repetitive assignments and banned them. For the same 100 tweets, we invited our research group again to perform this same experiment of relevance. Their assignments were collected along with the ones from Mechanical Turk. The results are shown in Table 18 showing an almost perfect agreement with Cohen's $\kappa = 0.854$.

6.3.2 GOLD STANDARD DATASET

Given this substantial agreement between the experts and the workers in regards to the proof of concept dataset, we can claim that Mechanical Turk can be used in estimating the content's time relevance and in turn to gauge the author's temporal intention after utilizing TIRM. The next step is to expand our dataset and collect a larger dataset, for training and testing, to utilize in the modeling process.

From the SNAP dataset of tweets we started by extracting a dataset of 20,000 tweets at random starting from June, 2009. For a social media post, in this case a tweet, we want to acquire as much data as possible about its existence such as content, age, dissemination, and size. Initially, we targeted the tweets which meet these criteria:

- The text is in the English language.

- Each has an embedded URI pointing to an external resource.

- The embedded URI has been shortened using Bitly.

- The embedded URIs point to unique resources.

- The linked resource is currently available on the live web.

- The resource has at least ten mementos in the public archives.

We chose tweets which have links because the scope of the study is focused on detecting intention in sharing resources in social media. Also, the shared resource provides extended context of the tweet, making the readers better grasp the message the author intended to convey. In a different light, the tweets can act as annotations to the linked resources. The reason behind choosing bitly shortened URIs is that their API provides invaluable information about the clicklog patterns, creation dates, rates of dissemination, and other information as will be described in the next section. Also bitly was popular on Twitter at the time of the dataset collection (2009). In 2010, Twitter released their own default URL shortener t.co, as mentioned earlier in Section 2.2, and the amount of tweets having bitly shortened URLs has decreased considerably. To ensure our ability to collect information related to the embedded resource, we only kept the linked resources that are currently available on the live web (HTTP response 200 OK) at the time of the analysis. Also we only kept the resources that are properly archived in the public archives with at least ten mementos each. Consequently, we extracted 5,937 unique instances to be utilized in the next stages.

To create the dataset that would be processed by Mechanical Turk workers, we randomly selected 1,124 instances from the previous dataset. This training dataset would be assigned to the workers in the same manner as the gold standard experiment described in Section 6.2. To have an insight of what the author was experiencing and reading upon the time of tweeting, we extracted the closest memento of the resource to the time of the tweet, using the Memento framework. For each URI, the closest memento recorded ranged from 3.07 minutes to 56.04 hours from the time of the tweet, averaging 25.79 hours. Figure 60 shows the difference in hours between t_{tweet} and the closest memento in the public archives denoted by $R_{closestMemento}$ for the top 1,000 instances. In the graph we account for the top 1,000, only not the whole 1,124, as a few URIs have around ten mementos which are spaced spanning a period of over ten years which makes the closest memento excessively far from the expected date. For the sake of simplicity we will consider the following approximation:

$$R_{closestMemento} \approx R_{tweet} \qquad (12)$$

This approximation shows that on average we can extract a snapshot of the state of the resource within a day from when the author saw it and tweeted about it. This time delta is in fact relative to the nature of the resource. In the case of continuously

Sorted time difference between the tweet and the closest memento
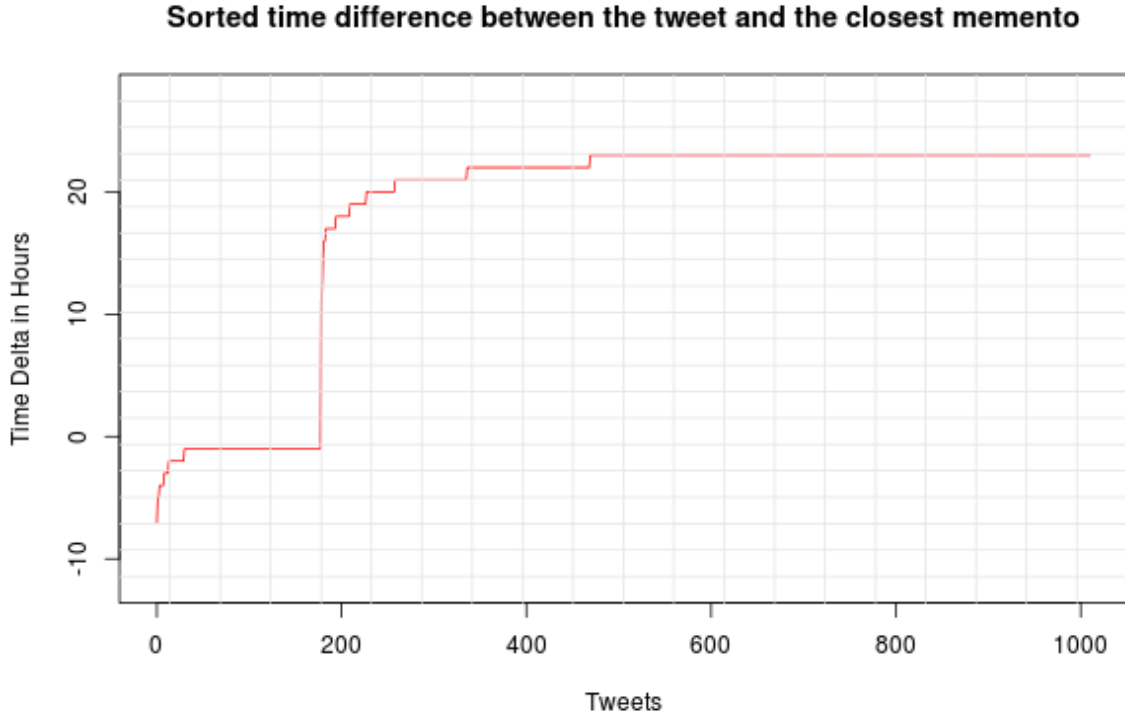
Figure 60. Sorted Time delta between tweeting time and the closest memento snapshot where the negative Y axis denotes existence prior to t_{tweet}

changing webpages, such as CNN.com, one day will not capture everything. Section 5.1 discusses the change in resources which are shared in the social web through time.

Along with the downloaded closest memento snapshot $R_{closestMemento}$, we downloaded a snapshot of the current state of the resource $R_{current}$. For the sake of simplicity as well, we consider another simplification:

$$R_{current} \equiv R_{click} \tag{13}$$

The agreement between Mechanical Turk workers in assigning relevancy to our training dataset of 1,124 tweets is shown in Table 19.

5 Turkers Agreeing (5-0 cuts)	589	52.40%
4 Turkers Agreeing (4-1 cuts)	309	27.49%
3 Turkers Agreeing (3-2 close call cuts)	226	20.11%
Relevant Assignments	929	82.65%
Non-Relevant Assignments	195	17.35%

Table 19. The distribution of voting outcomes from turkers for the 1,124 assignments

6.4 MEASURING CHANGE IN TIME

At this point we have successfully collected the Gold Standard Relevancy dataset with 1,124 instances that were assigned by turkers to belong to either the Relevant or Non-Relevant classes. The next step is to cover the other aspect of TIRM which is measuring the change in the resource from t_{tweet} to t_{click}. Following Equations 12 and 13, we have downloaded both versions of the resource $R_{closestMemento}$ and $R_{current}$ for each instance in the dataset.

To measure change, we used similarity measures in textual content (which is deeply studied and analyzed) in our calculation and utilized Equation 14 to calculate normalized change between versions.

$$\Delta Change = 1 - Similarity \tag{14}$$

$$Similarity = cos(R_{closestMemento}, R_{current}) \tag{15}$$

As discussed earlier in Section 3.3.2, there are several techniques to measure similarity; here we utilize cosine similarity. We first downloaded the rendered HTML content and since we were only focused on the textual change in content, we eliminated the boilerplate tags by utilizing the boilerplate removal from HTML pages and full text extraction algorithms by Kohlschütter et al. [208]. Kohlschütter released a Java implementation called Boilerpipe based on the algorithm[2]. We used python wrapper implemented by Misja Hoebe based on the original Java implementation[3]. Then we transformed the resulting text into a bag-of-words and in turn to word vectors and finally, we calculated the cosine similarity between the vectors

[2]https://code.google.com/p/boilerpipe/
[3]https://github.com/misja/python-boilerpipe

corresponding to each of the pairs of documents $R_{closestMemento}$ and $R_{current}$ as shown in Equation 15. This resulted in a normalized value of similarity with 0.0 denoting no similarity and 1.0 denoting identical content.

6.5 SUMMARY

In this chapter, we investigated the problem of the temporal inconsistency in social media and how it is related to the author's intention. This intention proved to be non-trivial to capture and gauge. Our Temporal Intention Relevancy Model (TIRM) successfully translated the problem of user intention to a less complicated problem of relevancy. We used Mechanical Turk to collect a gold standard data of user temporal intention and we verified the results by comparing the turkers' assignments to ones conducted by experts in the field and produced a near perfect agreement. After proving the validity of using Mechanical Turk in data gathering, we proceeded in collecting a dataset that was used in training the classifier.

The next step is to use TIRM and the gold standard dataset to create a classifier to assess relevancy and in turn model intention.

CHAPTER 7

MODELING INTENTION WITH RESPECT TO TIME

"Sometimes it's a little better to travel than to arrive"

— Robert M. Pirsig, *Zen and the Art of Motorcycle Maintenance: An Inquiry Into Values*

In Chapter 6, we collected the gold standard dataset using Mechanical Turk and tested its validity against expert opinions. The dataset collected contains tweets, which have embedded shortened URIs or bitlys linking to a shared web resource. Each one of the resources is currently live and adequately covered in the public web archives at the time of that experiment (December 2012). In this chapter we extend our analysis of intention to the next phase, as shown in Figure 61, which is modeling the intention. We analyze collections of features from social, archival, link, and textual aspects as shown in the following sections to train a model to identify human perception of relevance and map this modeling back to intention.
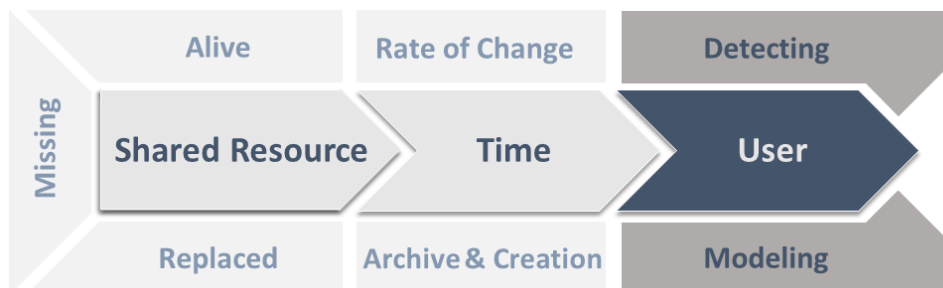
Figure 61. Modeling temporal intention

7.1 FEATURE EXTRACTION

To complement the training dataset we collected in the previous section (Table 19) from Mechanical Turk, we explore the different angles of sharing resources in

social media beyond the tweet. For each instance we have the original URI, the tweet textual content, the bitly URL, the timestamp of the tweet, and we downloaded both the current version of the resource $R_{current}$ along with the closest memento $R_{closestMemento}$ as described in Section 6.4. We continue by analyzing several aspects of the components of the problem and extract the corresponding features to each angle as follows.

7.1.1 LINK ANALYSIS

In the SNAP tweet dataset, out of the 476 million tweets in the dataset, 87 million contain bitly shortened URIs. The bitly API provides several parameters that we extracted like the total number of clicks, hourly clicklogs, creation dates, referring websites, referring countries, and other information could also be acquired.

The depth of the resource in the website is important as well. Surface web pages, as the main page or the index, are different in nature from the deep web pages. Figure 62 shows two sample webpages from CNN.com, one is a top level webpage (Figure 62a `http://www.cnn.com/world`), and the other is a deep level webpage (Figure 62b `http://www.cnn.com/2015/03/23/world/steve-mccurry-afghan-girl-photo/index.html`). The top level page changes on a regular basis corresponding to breaking news, while the deep level page tends to remain static, save for ads. This phenomenon is witnessed more often than not, so relying on this general notion that pages in the deep web are less likely to change as often as the root page, we need to calculate the estimated depth of the resource. Within each tweet, we expanded the resource's bitly to the original long URI and calculated the resource's depth by counting the number of backslashes in the URI. Also we compare the lengths of the shortened URl and the original one to calculate the reduction rate. Hand in hand with these extracted data points, we proceed to examine the dissemination trends of that resource.

7.1.2 SOCIAL MEDIA MINING

For each embedded resource in a tweet, we used Topsy.com's API to extract the total number of tweets that have been recorded linking to this resource. We extract the number of tweets from influential users in the Twitter-sphere as defined by Topsy (Figure 63). Finally, we downloaded the other tweets posted by different users linking to the same resource. The API permits a maximum of 500 tweets per resource. This collection of tweets surrounding each resource can benefit us in many
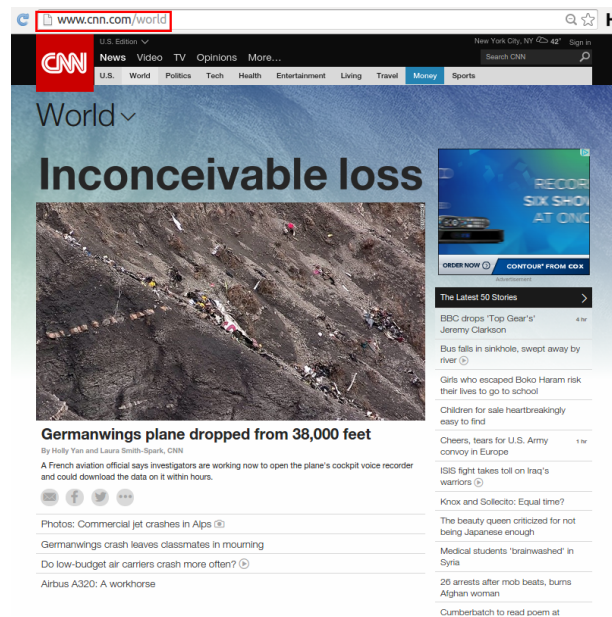
(a) Top level CNN page with depth = 1 http://www.cnn.com/world

(b) Deep level CNN page with depth = 6 http://www.cnn.com/2015/03/23/world/steve-mccurry-afghan-girl-photo/index.html

Figure 62. The top page will change more frequently than the bottom page

Figure 63.   Screenshot of Topsy's page of tweets linking to: http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html

aspects: providing extended tweet-context for the resource, showing us the social media dissemination pattern by plotting the tweet timestamps against the timeline, and finally, to let us examine how many of those tweets still exist and how many have been deleted.

To complete the picture, Facebook was mined as well for each of the resources in the tweets to extract the total number of shares, posts, likes, and clicks.

7.1.3 ARCHIVAL EXISTENCE

To investigate archival existence and coverage, we calculated how many total mementos in the aggregated public archives are available for the resource. We

also record how many archives hold at least a copy of the resource. As mentioned earlier, Figure 60 shows the distribution of the delta of time between closest archived memento $R_{closestMemento}$ and the tweet creation timestamp t_{tweet}. Negative values on the y-axis denote existence prior to t_{tweet}.

7.1.4 SENTIMENT ANALYSIS

To go beyond the tweet text, we utilized the NLTK libraries [223] for natural language text processing to extract the most prominent sentiment in the text. For each tweet we extracted the positive, negative and neutral sentiment probabilities. These three probabilities give us an insight on the emotional state of the author at t_{tweet}.

7.1.5 CONTENT SIMILARITY

In Section 6.4, we described how we measured the similarity between the different snapshots of the resource downloaded earlier at t_{tweet} and t_{click}. We downloaded the HTML, performed boilerplate removal and extracted the textual content. Next we transformed this textual content into vectors for each of the resource's R_{tweet} and R_{click} and then calculated the cosine similarity between them. It is also worth mentioning that using the boilerplate removal algorithm along with cosine similarity gave more significant features than raw HTML similarity with SimHash [108]. Furthermore, the collected tweets from Topsy.com's API associated to each resource have been accumulated in one document, giving it a social context. Section 4.3.1 describes in detail how we built this tweet document. Finally, we also transformed the tweet document into vector form to calculate its cosine similarity between R_{tweet} and R_{click}. The rationale behind this is to see if the textual "aboutness" of the resource has changed in social context with time.

7.1.6 ENTITY IDENTIFICATION

After analyzing hundreds of tweets from Twitter timeline, we noticed some interesting points. Celebrities are mentioned in abundance and have the largest number of followers. In fan tweets, most celebrities are mentioned by their first and last name unless they are known by only one, and finally most tweets about celebrities are in reaction or as a description to contemporaneous events related to the celebrity. In the fields of TV, cinema, performance arts, sports, and politics, millions of tweets

are posted daily about celebrities as a huge demographic of users use Twitter as a form of news feed. Given so, we wanted to analyze the effect of detecting celebrity-related tweets to intention and the possibility of using it as a feature. Wikipedia has published several lists of US, British, and Canadian actors and singers along with several lists of sports players and politicians in the English speaking world. We harvested, parsed, and indexed those lists. Finally, given an embedded resource, its corresponding URI, and all the tweets containing that URI from Topsy.com's API we test for the existence of celebrity entities in the collective tweets and record celebrity-relevance feature as true if a celebrity is present.

7.2 MODELING AND CLASSIFICATION

In the feature extraction phase we gathered several data points denoting context, dissemination, nature, archiving coverage, change, sentiment, and others. In this phase, we investigate which features have higher weights indicating importance in modeling and classifying temporal intention. We also investigate several well-known classifiers and their corresponding success rates.

In the first attempts to train the classifier and analyze the confusion matrix, we noticed the instances which were classified by Mechanical Turk workers as close calls (3-2 split) highly populated the false-positive and false-negative cells of the confusion matrix. These instances indicate a weak classification where one vote can deem the instance relevant or non-relevant. Thus to reduce the confusion, we eliminated these instances. From the 1,124 instances, we kept 898 where the agreement on relevancy was 4 to 1, or 5 total agreement, as shown in Table 20. Thus, the cutoff threshold in Equation 11 is increased $a \geq 0.8$.

Relevant Votes	807	89.87%
Non Relevant Votes	91	10.13%

Table 20. The distribution of voting outcomes from turkers after removing close-calls

Utilizing the sum of all the extracted features, we ran Weka's [224] different classifiers against the dataset. Subsequently, we train the model and test it using 10-fold cross validation. Tables 21 and 22 show the corresponding precision, recall

and F-measure of the Cost Sensitive classifier based on Random Forest, which out-performed the other classifiers yielding an 90.32% success in classification for our trained model.

	10-Fold Cross-Validation Testing
	Cost Sensitive Random Forest
Mean Absolute Error	0.15
Root Mean Squared Error	0.27
Kappa Statistic	0.39
Incorrectly Classified	**9.68**%
Correctly Classified	**90.32**%

Table 21. Results of 10-fold cross-validation against the best classifier along with the Precision, Recall and F-measure per class

Classifier	Precision	Recall	F-measure	Class
Cost Sensitive	0.93	0.96	0.95	**Relevant**
Random Forest	0.53	0.37	0.44	**Non-Relevant**
Weighted Average	**0.89**	**0.90**	**0.90**	

Table 22. Precision, Recall and F-measure per class

Rank	Feature	Gain Ratio
1	Existence of celebrities in tweets	0.149
2	Number of mementos	0.090
3	Tweet similarity with current page	0.071
4	Similarity: Current & past page	0.053
5	Similarity: Tweet & past page	0.044
6	Original URIs depth	0.032

Table 23. Classifier features ordered by significance resulting from Rank Search algorithm

The classifier processed 39 different features for each instance in the training dataset as shown in the top part of Table 26. The features were collected in the feature extraction phase explained earlier in Section 7.1. Following the training phase we needed to understand the effect of each feature in the process of modeling intention. This knowledge will help us in reducing the number of features required by the model to estimate the intention behind a given social post. We applied an attribute evaluator supervised algorithm based on the Ranker search method to rank the attributes or features accordingly. Analyzing the ranks, Table 23 shows the strongest six features and the order of significance in ranking the features used in classifying user temporal intention along with the information gain of each.

7.3 EVALUATION

The previous section indicates that modeling user intention via TIRM and using numerical, textual, and semantic features in a classifier is both feasible and accurate. In this section, we test the trained model against other tweet datasets.

7.3.1 EXTENDED DATASET

In Section 6.3.2 we extracted a dataset of 5,937 tweet-resource pair instances from which we extracted our training 1,124 instances training dataset. The remaining 4,813 instances formed a new testing dataset. For each instance in this dataset we extracted all the features analyzed in Section 7.1. Finally, this dataset was evaluated by the trained model to test the performance and usability yielding the results in Table 24.

7.3.2 HISTORICAL INTEGRITY OF TWEET COLLECTIONS

As described in Chapter 1, one of the main motives of our analysis of human intention is to maintain the historical integrity of social post collections. Specifically, in social posts related to historic events, preserving the consistency between the tweet and the linked resource is crucial. The link between the post and the resource is vulnerable to two kinds of threats: the loss of content itself (either the post or the linked resource) or the mismatch between the author's intention and what the reader is receiving (the resource is no longer intended by the author). In Section 4.1, we analyzed six datasets related to six different historic events and we evaluated how many of these resources are missing and how many are archived [204]. In this section,

we utilize our trained model in predicting the temporal intention and in turn, in estimating the amount of mismatched resources where the reader is probably not reading the first draft of history intended by the tweet's author.

To reiterate, the datasets from Section 4.1 cover the 2009-2012 events related to Michael Jackson's death, H1N1 virus outbreak, Iranian elections, President Obama's Nobel Peace Prize, and the Syrian uprising. Similarly to the extended testing dataset in Section 7.3.1, we extract all the necessary features for each instance in the dataset. We test our model with the five datasets and show the results in Table 24. For each dataset, we dereference the URIs again and record the response headers to assess the percentage alive (status 200 OK) and missing (Status 404 or Other). We started the experiments in September of 2012 and we recorded the percentage of missing resources in the 3,124 instances extended dataset. It is worth mentioning that after four months, we re-tested their existence and we noticed a loss of 3.23%, confirming the results from our previous work explained in Section 4.2.

Dataset	Status 200	Status 404 or Other	Relevant Percentage	NonRelevant Percentage
Extended 3,124 instances	96.77%	3.23%	96.74%	3.26%
MJ's Death	57.54%	42.46%	93.24%	6.76%
H1N1 Outbreak	8.96%	91.04%	97.48%	2.52%
Iran Elections	68.21%	31.79%	94.69%	5.31%
Obama's Nobel	62.86%	37.14%	93.89%	6.11%
Syrian Uprising	80.80%	19.20%	70.26%	29.75%

Table 24. Results of testing the extended dataset & the historic datasets in classifying relevancy along with the live percentage, and percentage missing of the resources

7.3.3 EVALUATING TIRM

After examining the relevancy of the datasets using our developed relevancy classifier, we now use our TIRM mapping scheme in transforming the results into the intention space. The classifier was trained to be conservative in handling the Non-Relevant categorization, which means classifying Non-Relevancy false negatives is more tolerated than false positives (i.e., the classifier only states a resource is

		Relevant	Not Relevant
Changed	MJ	41%	3%
	Obama	42%	2%
	Syria	44%	25%
	Iran	49%	2%
	H1N1	6%	0%
	Extended	53%	2%
Not Changed	MJ	52%	4%
	Obama	51%	5%
	Syria	26%	5%
	Iran	46%	3%
	H1N1	91%	3%
	Extended	43%	2%

Table 25. TIRM classification for the six historical data sets

non-relevant only if it was highly confident of this estimation). Another point worth mentioning is that, for our training, we used the resources that are currently available on the live web and 404 resources were not included. Table 25 show the percentages in each of the six datasets per each class of the TIRM model after mapping relevancy to the similarity threshold of 70%. Taking the dataset of Michael Jackson's death for example, nearly 3% of the dataset is still accessible but is no longer relevant. It is worth noting that the results in the first quadrant of Table 25 are over-reported. Due to the sparsity of the archives, this over reporting is essential to avoid false negatives. As shown in Figure 60, the average time delta between sharing and the closest archived version is considerably large (26 hours), in some cases the resource will keep on changing then stops after a couple of hours and stay static. Tightening the bounds in the same figure by more frequent archiving will lead to a large improvement in our model.

7.4 ENHANCING TIRM

To this point we were able to successfully model temporal intention by decomposing it into two simpler tasks of *content relevance* and *change*.

While ranking the 39 features extracted earlier in Section 7.1 with respect to information gain, we made several intriguing observations. First, the simple detection of celebrities in the tweet was ranked atop of the list. A possible explanation is that

when users discuss topics or events related to a celebrity they most likely target contemporaneous events or breaking news, like scandals, rather than long-term events. This observation of high information gain corresponding to entity recognition (in this case celebrities) highlights the need of incorporating further linguistic analysis in our model.

In TIRM, the first stage is to identify if the resource is still relevant to the tweet or not, then we measure how the current state of the resource has changed or not from the archived version at the time of the tweet. We noticed that the classification was greatly biased towards the "Relevant" class, which also highlighted the need to enhance the dataset and remove that bias by balancing it. Finally, we observed that we performed a word-based textual comparison in order to calculate the similarity between the tweet and the resource, which proved lacking since the tweet is limited to 140 characters while the resource could span thousands of characters. This highlights the need to find a better similarity measure based on the semantic similarity rather than simple term overlap.

To address these observations, we developed a three-staged approach in enhancing the prior model on the following aspects:

- Linguistic analysis of the tweet.

- Semantic similarity measure instead of a lexical similarity.

- Fixing the training dataset and remove the inherent bias towards the "Relevant" class.

Following the enhancement of the model, we want to estimate the confidence of this probabilistic classification. Beyond mapping intention to a class, we need to quantify this intention in order to measure it. With the model in its primitive phase, we were able to detect and classify the temporal intention to either *Current* or *Past*, but how certain are we of this intention? We propose a formulation to intention based on the relevance measure from the classifier and the change measure obtained by calculating similarity between the resource's versions. We call this formulation the *Intention Strength Measure*.

As discussed above, we utilize TIRM as shown previously in Table 17 and enhance its performance and improve its accuracy. In this enhancement stage we utilize the same dataset of 1,124 instances.

7.4.1 LINGUISTIC FEATURE ANALYSIS

Previously, 39 different features were extracted from the tweet-resource pair in regards to similarity, URL structure, social and archival existence. The results were promising but we needed a deeper analysis and understanding of the linguistic properties of the tweet-resource pair. At this stage we enhanced the model by extending those features starting with a deeper linguistic analysis of the tweet, and the resource at both t_{tweet} and $t_{current}$.

Tweet structural analysis

After removing the URI of the linked resource we checked remaining tweet text for the existence of user mentions, hashtags, question marks "?" (indicating a question tweet), and exclamation marks "!" (indicating an expression of strong feelings). Furthermore, we utilized regular expressions, adopted from Ritter et al.'s work, in detecting emoticons in the tweets [225]. We deduced that along with the extracted sentiment from the prior experiment, we would be able to capture the emotional state the author. Finally, we also checked if the tweet was a re-tweet. These simple features proved to be highly effective, as six of which are present among the top 13 ranked features in information gain of the retrained model (Table 26).

POS tagging and Named Entity Extraction

In the prior TIRM, we harvested Wikipedia for lists of artists, actors, and singers from the English speaking world to use in detecting the existence of celebrities in the tweets. This feature proved to be highly valuable due to its corresponding high information gain. This observation led us to believe we need to further investigate named entities in tweets.

In tweet analysis, due to the 140 character limit and corresponding lack of context and the informality in writing, tasks like part-of-speech (POS) tagging, sentence chunking, and named entity recognition are quite challenging. Ritter et al. developed a distantly supervised approach that is tailored for tweet based analysis overcoming those challenges [225]. We adopted their labeled LDA-based POS tagger and chunker, which have performed effectively against standard POS taggers on tweet datasets. Ten different types of entities are defined by Ritter's tagger as

#	Feature Name	Type	Extraction Method	Availability	Gain	Rank	Min	
1	ShortURLLen	Structural	Analyzing resource's URL	At t_{tweet}	0.1709	4	✓	
2	NumArchives	Archival	Analyzing resource's timemap	After Archival	0.1663	5	✓	
3	URLDepth	Structural	Analyzing resource's URL	At t_{tweet}	0.1569	10	✓	
4	CelebInTwts	Linguistic	Mining Wiki+Text analysis	At t_{tweet}	0.1203	11	✓	
5	CelebInTwt	Linguistic	Mining Wiki+Text analysis	At t_{tweet}	0.0917	22	✓	
6	CosTwtPast	Similarity	Cosine Similarity+BoilerPlt	After Archival	0.0877	23	✓	
7	CosCurTwt	Similarity	Cosine Similarity+BoilerPlt	At t_{tweet}	0.0864	24	✓	
8	CosCurPast	Similarity	Cosine Similarity+BoilerPlt	After Archival	0.0862	25	✓	
9	CelebPctInTwt	Linguistic	Mining Wiki+Text analysis	At t_{tweet}	0.0861	26	✓	
10	TwtSimCur	Similarity	Similarity+BoilerPlt	At t_{tweet}	0.0846	27	✓	
11	URLLen	Structural	Analyzing resource's URL	At t_{tweet}	0.0846	286		The original TIRM Model with 39 Features
12	ReductionRate	Structural	Analyzing resource's URL	At t_{tweet}	0.0845	29		
13	CelebPctInTwts	Linguistic	Mining Wiki+Text analysis	After being retweeted	0.0835	30		
14	InfluTwtsCount	Social	Mining Topsy API	After being retweeted	0.0835	31		
15	SimhashCurPast	Similarity	Simhash Similarity+BoilerPlt	After Archival	0.0799	33		
16	MementoCount	Archival	Analyzing resource's timemap	After Archival	0.0774	34		
17	FBClicks	Social	Mining FB API	After being posted on FB	0.074	35		
18	CosCurTwts	Similarity	Cosine Similarity+BoilerPlt	After being retweeted	0.0695	36		
19	FBLikes	Social	Mining FB API	After being posted on FB	0.0689	37		
20	FBComments	Social	Mining FB API	After being posted on FB	0.0668	38		
21	TwtLen	Structural	Text analysis	At t_{tweet}	0.0662	39		
22	CosTwtsPast	Similarity	Cosine Similarity+BoilerPlt	After Archival+retweeted	0.0569	41		
23	SimhashCurTwts	Similarity	Simhash Similarity+BoilerPlt	After being retweeted	0.0569	42		
24	FBShares	Social	Mining FB API	After being posted on FB	0.0538	44		
25	InitContentLen	Structural	Mining Bitly API	After being Bitly Shortened	0.0481	46		
26	NeuSentiment	Linguistic	NLTK Sentiment Analysis	At t_{tweet}	0.048	47		
27	TwtSimPast	Similarity	Similarity+BoilerPlt	After Archival	0.0475	48		
28	BitlyClicks	Social	Mining Bitly API	After being Bitly Shortened	0.0463	49		
29	SimhashCurTwt	Similarity	Simhash Similarity+BoilerPlt	At t_{tweet}	0.0438	52		
30	CloseMemTime	Archival	Analyzing resource's timemap	After Archival	0.0434	53		
31	SimhashTwtpast	Similarity	Simhash Similarity+BoilerPlt	After Archival	0.0411	55		
32	PastCurSim	Similarity	Similarity+BoilerPlt	After Archival	0.0376	56		
33	PosSentiment	Linguistic	NLTK Sentiment Analysis	At t_{tweet}	0.0356	57		
34	SimhashTwtsPast	Similarity	Simhash Similarity+BoilerPlt	After Archival+retweeted	0.0353	58		
35	TwtsSimCur	Similarity	Similarity+BoilerPlt	At t_{tweet}	0.0351	59		
36	RetrievedTwts	Social	Mining Topsy API	After being retweeted	0.0233	60		
37	NegSentiment	Linguistic	NLTK Sentiment Analysis	At t_{tweet}	0.0215	62		
38	TotalTwtCount	Social	Mining Topsy API	After being retweeted	0.0202	63		
39	TwtsSimPast	Similarity	Similarity+BoilerPlt	After Archival	0	65		
40	UserMention	Linguistic	Text analysis	At t_{tweet}	0.2254	1	✓	
41	IsRetweet	Linguistic	Text analysis	At t_{tweet}	0.2015	2	✓	
42	Has!	Linguistic	Text analysis	At t_{tweet}	0.1845	3	✓	
43	GEO-LOC	Linguistic	Named Entity Extraction	At t_{tweet}	0.1653	6	✓	
44	Has?	Linguistic	Text analysis	At t_{tweet}	0.1643	7	✓	The enhancing extended features
45	PERSON	Linguistic	Named Entity Extraction	At t_{tweet}	0.1612	8	✓	
46	HashtagCount	Linguistic	Counting Hashtags	At t_{tweet}	0.1602	9	✓	
47	COMPANY	Linguistic	Named Entity Extraction	At t_{tweet}	0.1186	12	✓	
48	HasEmoticon	Linguistic	Text analysis	At t_{tweet}	0.1106	13	✓	
49	MOVIE	Linguistic	Named Entity Extraction	At t_{tweet}	0.1085	14	✓	
50	TVSHOW	Linguistic	Named Entity Extraction	At t_{tweet}	0.1065	15	✓	
51	OTHER	Linguistic	Named Entity Extraction	At t_{tweet}	0.1056	16	✓	
52	BAND	Linguistic	Named Entity Extraction	At t_{tweet}	0.1016	17	✓	
53	SPORTSTEAM	Linguistic	Named Entity Extraction	At t_{tweet}	0.0985	18	✓	
54	LDATwtsSimCur	Similarity	LDA Similarity+BoilerPlt	After being retweeted	0.0945	19	✓	
55	PRODUCT	Linguistic	Named Entity Extraction	At t_{tweet}	0.0922	20	✓	
56	LSATwtSimCur	Similarity	LSA Similarity+BoilerPlt	At t_{tweet}	0.092	21	✓	
57	LSATwtsSimCur	Similarity	LSA Similarity+BoilerPlt	After being retweeted	0.0819	32		
58	LSATwtSimPast	Similarity	LSA Similarity+BoilerPlt	After Archival	0.0591	40		
59	LSATwtsSimPast	Similarity	LSA Similarity+BoilerPlt	After Archival+retweeted	0.0548	43		
60	LDATwtSimCur	Similarity	LDA Similarity+BoilerPlt	At t_{tweet}	0.0522	45		
61	TweetClass	Linguistic	LDA Tweet Classification	At t_{tweet}	0.0453	50		
62	LDATwtSimPast	Similarity	LDA Similarity+BoilerPlt	After Archival	0.0452	51		
63	LDATwtsSimPast	Similarity	LDA Similarity+BoilerPlt	After Archival+retweeted	0.0429	54		
64	Tense	Linguistic	POS tagging	At t_{tweet}	0.0223	61		
65	FACILITY	Linguistic	Named Entity Extraction	At t_{tweet}	0	64		

Table 26. All TIRM, Enhanced TIRM, and Minimized TIRM, features ranked by Information Gain Ratio. **Key:** *FB=Facebook, Twt=Tweet, Sim=Similarity, Cur=Current, Len=Length, Celeb=Celebrities, Pct=Percent, Init=Initial, Pos=Positive, Neg=Negative, Neu=Neutral*

shown in Table 27, along with the number of identified entities in each class across the training dataset of 1,124 instances. Furthermore, with the extracted POS tags and chunks, we were able to determine if the most prominent tense in a tweet is present or past and used it as a feature too. The rationale behind this analysis is to also identify the intention of the author in discussing contemporaneous events or past ones.

Entity Type	Instance Count
Person	233
Geo-Location	81
TV Show	18
Movie	37
Facility	19
Company	115
Product	42
Sports Team	10
Band	62
Other	96
Tweets with Named Entities	543
Tweets without Named Entities	581

Table 27. Named entities instances in the dataset

Tweet Classification

Users tweet to convey an opinion, update a status, ask for information, express sarcasm, spread jokes, and many other reasons [13]. In our search for the author's temporal intention we utilized Wang et al.'s work in classifying tweeting motive [226]. We adopted the first level of their two-tiered classification: Opinion, Update, Interaction, Fact, Deals, News, and Others. Furthermore, and for the sake of simplicity, we utilized only the largest classes of *Opinion, Update, Interaction, Others*, which collectively comprised 94% of the instances in Wang et al.'s dataset.

As shown in Table 28, for class *Interaction*, the *Relevance* class is significantly higher than the other, while in class *Opinion*, the instances are more biased towards the Non-Relevant. This indicates the relation between tweet class and relevance; thus we use it as a feature.

	Interaction	Update	Opinion
Relevant	69.67%	59.28%	36.99%
Non-Relevant	30.33%	40.72%	63.01%

Table 28. Tweet classification across relevancy classes

7.4.2 SEMANTIC SIMILARITY ANALYSIS USING LATENT TOPIC MODELING

In the prior TIRM, similarity measures were based on word overlap either by using SimHash or cosine similarity. We were faced by two major shortcomings in regards to the resource and the corresponding tweet. First, using Simhash and cosine similarity techniques proved to be lacking upon calculating the similarity between a tweet (140 characters) and a resource, which could be virtually unlimited in size. Second, between two versions of a resource, a change in the HTML design could be interpreted as a low similarity, while in fact the content itself remained unchanged. In our experiment, we attempted to overcome the latter problem by introducing a boiler plate removal algorithm to remove the effect of change in styling and extract the main content.

To address the former we employed topic detection, as we would consider a tweet and a resource to be similar if they were both mentioning the same topic or discussing the same point. Thus, we measure similarity based on collective semantics or "aboutness" of the pair rather than textual overlap.

We use both latent semantic analysis (LSA) [227] and latent Dirichlet allocation (LDA) [228] in calculating the similarities between the tweet-resource and resource-resource pairs accordingly. We considered both techniques as LSA (or interchangeably called LSI for Latent Semantic Indexing), which is much faster to train, while LDA has higher accuracy. We also considered utilizing Twitter-based LDA models from the works of Mehrotra et al. [229] and Zhao et al. [230], which are more fitted to handle tweeted textual content with its embedded hashtags. Since we were not performing topic modeling on tweets only and we are calculating similarities between the tweet and the resource, which is written formally than tweets in most cases, traditional LDA-LSI models trained on a diverse corpus like Wikipedia seemed more suitable. Furthermore, we calculated the similarities between the resource versions (R_{tweet} and R_{click}) and the tweet.

To prepare these models we utilize the Wikipedia Corpus in extracting the topics

and features. We downloaded 4,295,020 documents spanning the English Wikipedia documents in January 2014[1]. We chose Wikipedia for training, as it spans a wide variety of topics. Next we built the LDA and LSA models with 100,000 features, 672,235,199 non-zero entries in the sparse TF-IDF matrix. The LDA model in this case is an online learning LDA model developed by Hoffman et al. [231]. We collect R_{click}, R_{tweet}, and the tweet and convert each to latent vector space, and using the model we calculate the cosine similarity. The result is a number ranging from 0.0 (no similarity) and 1.0 (identical). Gensim by Řehůřek et al. was used in our LDA and LSA modeling and similarity calculations [232].

	10-Fold Cross-Validation Testing	
Model	**TIRM**	**Enhanced TIRM**
Mean Absolute Error	0.22	0.20
Relative Absolute Error	75.77	39.69
Kappa Statistic	0.31	0.81
Incorrectly Classified	15.12%	9.73%
Correctly Classified	84.88%	90.27%

Table 29. Results of 10-fold cross-validation for TIRM and after the three-staged enhancement process

7.4.3 DATASET BALANCING

From the prior experiment explained in Section 6.3.2, the dataset used in training and cross validation was collected using five different Mechanical Turk voters for 1,124 instances. The instances were classified by the majority of voters as Relevant and Non-Relevant classes. Unfortunately, but yet matching intuition, the dataset collected is biased towards Relevancy (with 930 Relevant vs. 194 Non-Relevant). This undersampling of the class Non-Relevant is causing the trained model to be more aggressive towards the Relevant class as shown in the class-based recall, precision and F-measure in Table 30.

The problem of imbalanced training datasets in classification is a well-known problem. In a multitude of cases, one of the classes is significantly lower in training

---

[1]http://download.wikimedia.org/enwiki/

	Precision	Recall	F-measure
TIRM			
Relevant	0.863	0.971	0.914
Non-Relevant	0.654	0.263	0.375
Weighted Avg.	0.827	0.849	0.821
Enhanced TIRM			
Relevant	0.880	0.932	0.905
Non-Relevant	0.928	0.873	0.900
Weighted Avg.	0.904	0.903	0.903
Minimized TIRM			
Relevant	0.849	0.939	0.892
Non-Relevant	0.932	0.834	0.880
Weighted Avg.	0.890	0.886	0.886

Table 30. Results from the TIRM, TIRM after enhancement, and TIRM after minimization with Random Forest Classifier

points than the other class(es). This causes the classifier to be overly sensitive towards one class than the other. In our analysis, the Relevant class is almost five times larger than the Non-Relevant class, resulting in a reduced precision and recall in the minor class. A possible solution to this problem is to undersample the major class (Relevant) to be nearly the same size of the minor class (Non-Relevant). This approach has a downside, as we purposely disposed of good data points that could enhance the classifier. Also, it gravely reduces the size of the training dataset for the collective classes.

Another approach is the Synthetic Minority Over-sampling Technique (SMOTE) introduced by Chawla et al. [233]. By synthesizing balancing datapoints via over-sampling the minor class in the dataset and utilizing the k-nearest neighbors algorithm, they were able to enrich the training dataset iteratively by oversampling the minor class until the two classes were close in size. Their technique proved to achieve better classifier performance (in ROC space) than undersampling the major class. Given this, we utilized SMOTE with five nearest neighbors in balancing our Relevant-NonRelevant dataset iteratively, and then we randomized the dataset uniformly.

7.4.4 FEATURE MINIMIZATION

To this point, we have collected 65 different features (39 original + 26 new) to train TIRM. Due to the associated high cost of calculating all the features, we investigate the effect of feature minimization on the trained classifier.

For each feature, there are two important factors: cost (computational power and time) and effectiveness (information gain ratio). We will assume a uniform cost and optimize in regards to information gain. We use ranker algorithm in extracting the top 25 features (as shown in Table 26) in terms of information gain to retrain TIRM. Table 30 shows the ~60% reduced TIRM classifier has a performance reduction of about 2%.

7.5 INTENTION STRENGTH

To indicate the intention class, we use the resulting relevance from the model along with change in TIRM (as illustrated in Table 17). This mapping model is effective, but unfortunately, although we can deduce the intention class (being past or current), there is no quantification of this intention strength. To overcome this, we devise a formulation of calculating the intention strength in terms of change and relevance as follows.

For each resource r, the similarity $\sigma_{past-current}$ is calculated using LDA similarity illustrated earlier between the two versions, R_{tweet} and R_{click}. The $\delta_{past-current}$ change is calculated in Equation 16.

$$\delta_{past-current}(r) = 1 - \sigma_{past-current}(r) \tag{16}$$

From the classifier we extract the relevance measure $\rho(r)$ ranging from 0.0-1.0, with 0.0 being completely Non-Relevant and 1.0 being completely Relevant. Referring back to the TIRM model Table 17 we define the intention class $\chi(r)$ in terms of change $\delta(r)$ and relevance $\rho(r)$ as follows:

$$\chi(r) = \begin{cases} \text{Current}, & \text{if } \rho(r) > 0.5 \ \& \ \delta(r) > 0.5 \\ \text{Past}, & \text{if } \begin{cases} \rho(r) < 0.5 \ \& \ \delta(r) > 0.5 \\ \rho(r) > 0.5 \ \& \ \delta(r) < 0.5 \end{cases} \\ \text{Unknown}, & \text{otherwise} \end{cases} \tag{17}$$
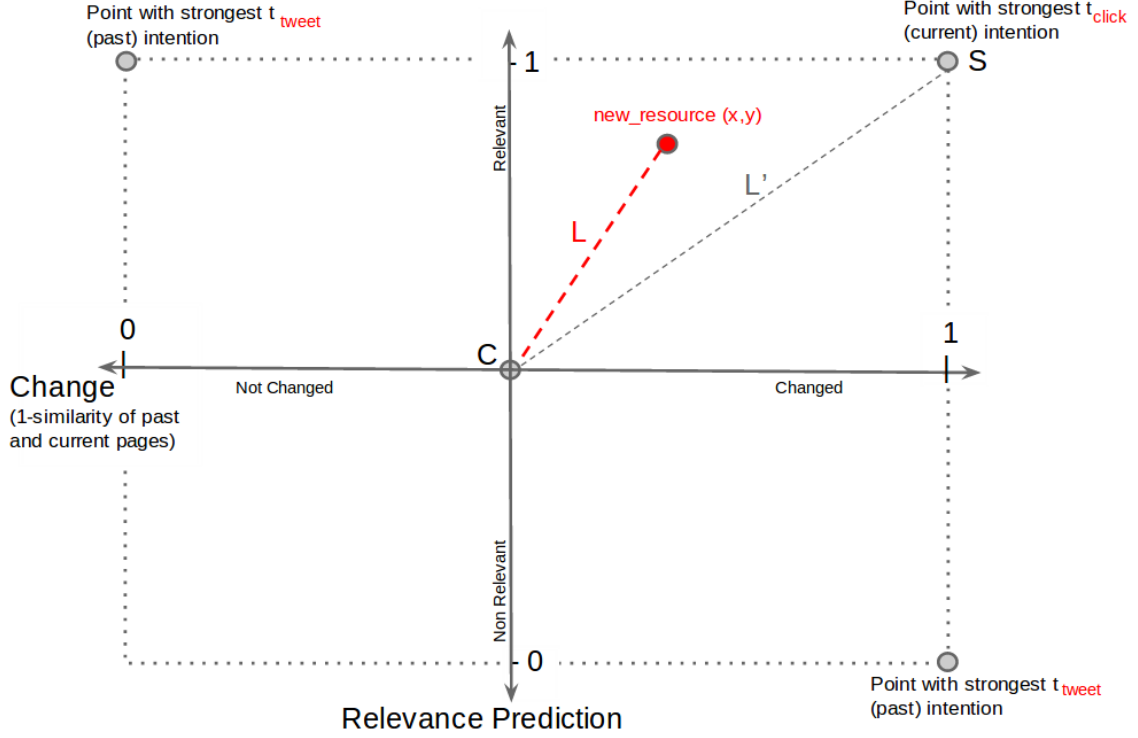
Figure 64. Intention Strength mapping

After identifying the intention class $\chi(r)$, we calculate the intention magnitude or strength $|\chi(r)|$. From Figure 64 we can deduce that the point $(\rho(r_s),\ \delta(r_s)) = (1.0,1.0)$ means it is most relevant and completely changed, which indicates the strongest "decided current intention" or $|\chi(r_s)| = 1.0$.

Point $(\rho(r_c),\ \delta(r_c)) = (0.5,0.5)$ is considered the point of confusion, as it illustrates peak uncertainty of intention, or $|\chi(r_c)| = 0.0$. The further the new resource $(\rho(r),\ \delta(r)) = $ (x,y) is from the point of confusion the stronger the intention certainty is. The furthest distance is the distance from the confusion point $(\rho(r_c),\ \delta(r_c)) = (0.5,0.5)$ to certainty point $(\rho(r_s),\ \delta(rs)) = (1.0,1.0)$. This Euclidean distance S will be used for normalization.

So to calculate $|\chi(r)|$ for the new resource $(\rho(r),\ \delta(r)) = $ (x,y) we follow Equation 18.

$$|\chi(r)| = \frac{L}{L'} = \frac{\sqrt{(\rho(r) - \rho(r_c))^2 + (\delta(r) - \delta(r_c))^2}}{\sqrt{(\rho(r_s) - \rho(r_c))^2 + (\delta(r_s) - \delta(r_c))^2}} \qquad (18)$$

Or to simplify:

$$|\chi(r)| = \frac{\sqrt{(\rho(r) - \frac{1}{2})^2 + (\delta(r) - \frac{1}{2})^2}}{\sqrt{(1 - \frac{1}{2})^2 + (1 - \frac{1}{2})^2}} \quad (19)$$

$$|\chi(r)| = \sqrt{2[(\rho(r) - \frac{1}{2})^2 + (\delta(r) - \frac{1}{2})^2]} \quad (20)$$
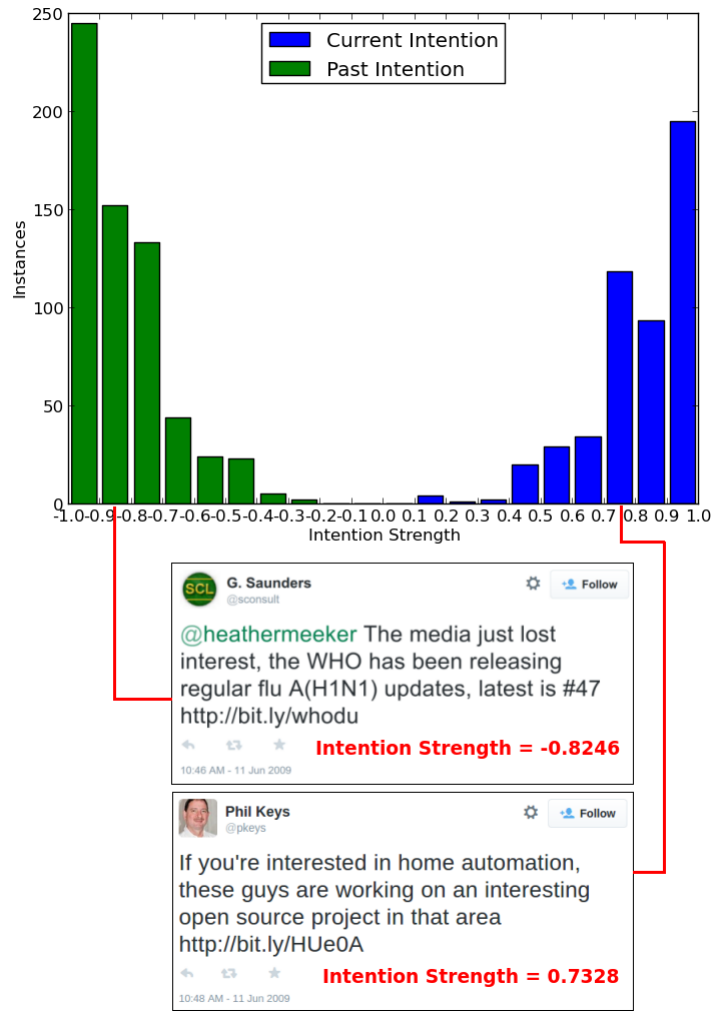
Figure 65. Histogram of the 1,124 instances in each intention strength bin with two example tweets

Finally by merging the intention class $\chi(r)$ and the intention strength $|\chi(r)|$ we get:

$$|\chi(r)| = \begin{cases} |\chi(r)| & if \;\; \chi(r) = Current \\ -|\chi(r)| & if \;\; \chi(r) = Past \\ Undefined & if \;\; \chi(r) = Unknown \end{cases} \qquad (21)$$
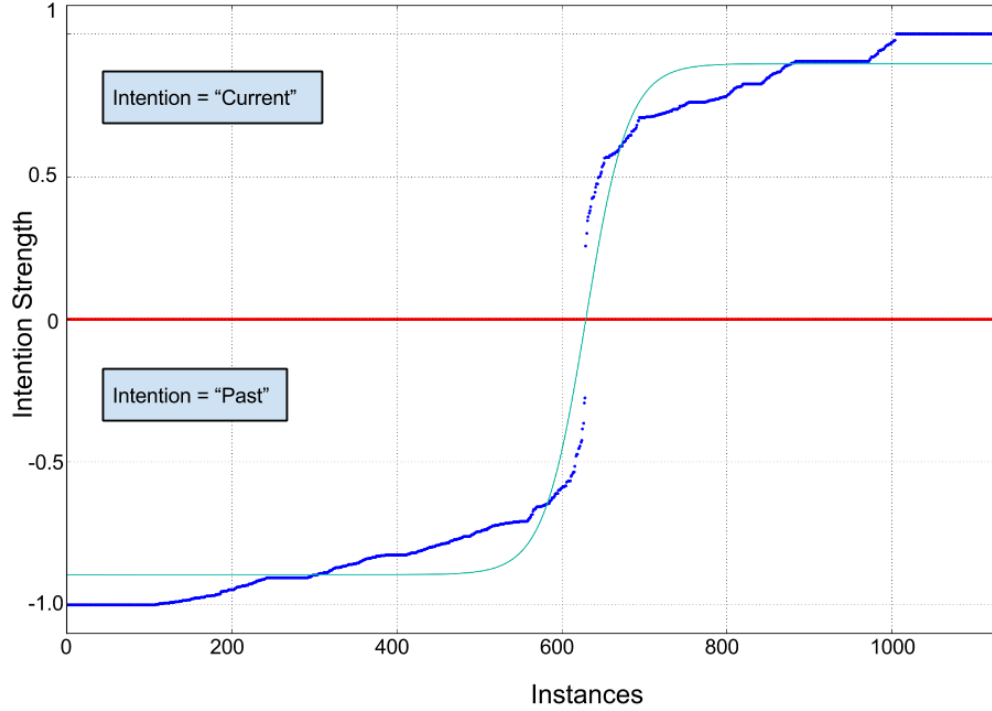
Figure 66. Intention strength across all 1,124 instances

Equation 21 summarizes $|\chi(r)|$ to be a value ranging from -1.0 to 1.0, with -1.0 being the strongest *Past* intention and 1.0 being the strongest *Current* intention. For the 1,124 instances in the dataset we calculate the corresponding intention strengths $|\chi(r_{1-1,124})|$. Figure 65 shows a histogram of the instances in each intention strength bin ranging from -1.0 to 1.0 and Figure 66 shows the sorted instances in terms of intention strength.

7.6 SUMMARY

In this chapter, we continued analyzing the problem of temporal intention in sharing resources in social media. We extracted several numerical, textual, and semantic features and incorporated them in the training dataset. The trained model

is then evaluated against an extended larger dataset and the datasets from our previous work regarding social posts from different six historical events in the period from 2009-2012. For the shared resources, we found temporal inconsistency to range from <1% to 25%, depending on the dataset. TIRM enabled us to detect and classify relevance and map it along with the resource's change to extract the intention class of the tweet in relation to the linked resource in it. We enhanced the model and addressed the shortcomings in regards to linguistic features analysis, balancing the training dataset, and finally used latent semantics in measuring similarity instead of merely textual resemblance. With these three stages, we were able to enhance the model considerably, especially in the Non-Relevant class, with a 0.5 improvement in F-measure and a 6% increase in total classification from the prior model upon utilizing a Random Forest-based classifier.

Finally, we formulated a method to quantify this temporal intention based on the enhanced model. Merging the new semantic change measure and the relevance prediction from the enhanced classifier, we produced a normalized quantifiable intention strength measure ranging from -1.0 to 1.0 (past to current intention, respectively).

CHAPTER 8

THE ROAD TO TEMPORAL INTENTION PREDICTION

"Whoever wishes to foresee the future must consult the past; for human events ever resemble those of preceding times. This arises from the fact that they are produced by men who ever have been, and ever shall be, animated by the same passions, and thus they necessarily have the same results" — Niccolò Machiavelli

At this point, we are able to calculate the intention (t_{tweet} or t_{click}) and the intention strength at the current time, given the tweet, $R_{current}$, and $R_{closestMemento}$. The next logical question was: Did the intention strength through the life span of the resource between t_{tweet} and t_{click} change at one point during these three and half years?

Answering this question will put us on track of answering the ultimate question of this chapter: Would the study of how intention strength changes through time allow intention prediction at t_{tweet}? This prediction as shown in Figure 67 is the third and final stage of the user behavioral analysis and the culmination of our user temporal intention analysis.
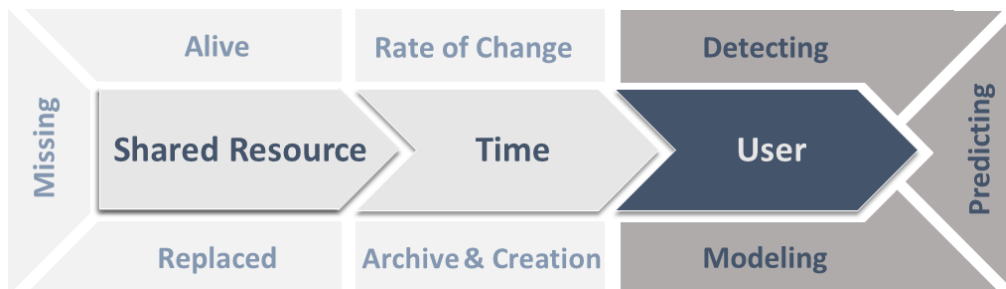
Figure 67. Predicting user's temporal intention

Consider the tweets shown in Figure 68. In Figure 68a we can see that the author's intention is for a specific information resource and thus the intention is t_{tweet}. In Figure 68b the author wants the reader to see the latest information, so the intention is t_{click}. In this chapter, we build a predictive model which can effectively differentiate between each intention class at tweet-authoring time (t_{tweet}). The ability to differentiate the intention in real-time can be used to push a copy of the linked resource into a web archive (e.g., webcitation.org, archive.today, archive.org) at t_{tweet} so the link is to an archived version instead of a web version, thus ensuring what readers see is consistent with the author's intention.
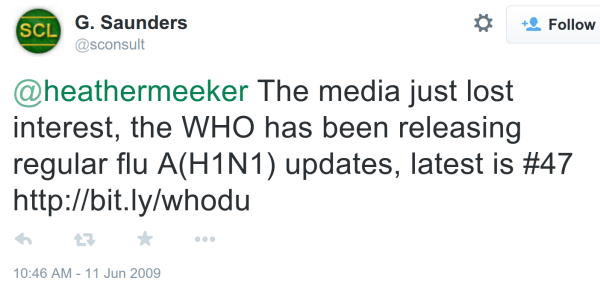
(a) The intention is towards the past version
at t_{tweet}

(b) The intention is towards the latest version
at t_{click}

Figure 68. Tweet examples for different intention classes

8.1 INTENTION AS A FUNCTION OF TIME

To recap, we examine two points in the life of a tweet as described earlier: 1) t_{tweet} when the author of the tweet posted it, 2) t_{click} when the reader clicks on the

link to examine the resource at current time. Table 17 shows that if the resource has changed and no longer relevant, then the intention is for the past (e.g., in Figure 68a the author intends for readers to access the WHO page as it was at t_{tweet}), while if the resource changed but still relevant, then the intention is for the current (e.g., in Figure 68b the author intends for readers to access latest news page as it will be at t_{click}). The model was trained using 65 different social, archival, contextual, and textual features extracted at t_{click}.

To recap on the modeling experiment in Chapter 7, from the SNAP dataset we extracted 1,124 tweets, we trained our classifier, and the *current* snapshots were captured in January 2013 after about three and half years from t_{tweet}. To get the past version of the resource, we extracted the closest memento ($R_{closestMemento}$) to the time of posting the tweet t_{tweet}. For the sake of simplicity, we assumed these time deltas are negligible and $t_{closest\_memento} \approx t_{tweet}$. Following the same paradigm we extracted ten mementos from the period between t_{tweet} and t_{click}:

$$t_{snapshot}(i) = \begin{cases} t_{tweet} & for \quad i = 0 \\ t_{memento}(i) & \forall i = 1...10 \\ t_{click} & for \quad i = 11 \end{cases} \tag{22}$$
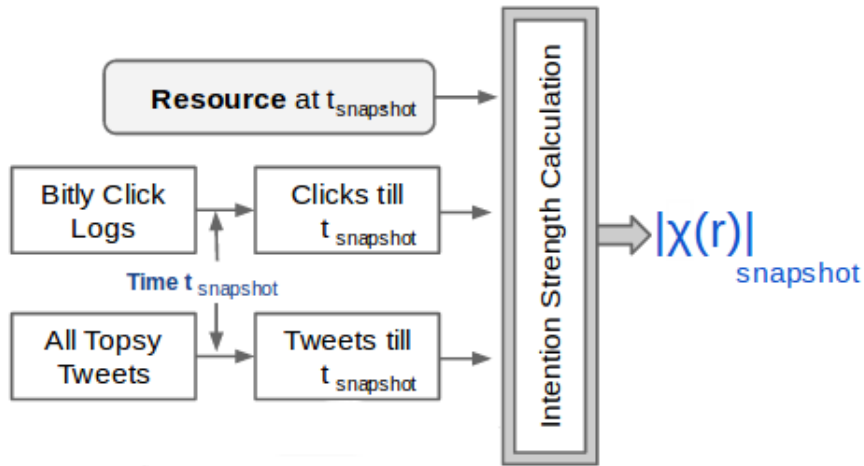
Figure 69. Intention Strength calculation per snapshot

Where $i = 0$ means the first snapshot which is at time of the tweet and $i = 11$ means the last snapshot at the current end time of the experiment. The ten

downloaded mementos are at $i = 1...10$.

The next step was to calculate the intention strength at each of those 12 points in time. Since we need to simulate the state at each time $t_{snapshot}(i)$ we need to download the state of the resource, get the Bitly clicklogs and the summation of the posted tweets up to this time. We mined the Bitly API to extract the clicks count to that moment $t_{snapshot}(i)$. We extracted the tweets posted until $t_{snapshot}(i)$ from Topsy.com API. This is another rationale behind using Topsy API instead of the Twitter API as the latter does not enable searching further than the indexing period (two weeks). This was true as of 2013, but in November 2014, Twitter enabled its new search index and permitted users to search for any tweet ever tweeted [234]. Furthermore, we calculated all the applicable features for each snapshot as shown in Figure 69. Finally, using our prior trained model and the strength formulation we calculated $|\chi(r_i)|$ for each snapshot and plotted them across time, as shown in examples in Figure 70.

8.2 PREDICTING TEMPORAL INTENTION AT TWEET TIME

In Figure 70, the blue points indicate the intention strength at this point in time. We noticed a steady behavior with respect to time in some cases and a changing behavior in others. This matches our intuition that users intended for the readers to see the version at t_{click} for the first short period of time, but upon changing and updating of the resource the intention deviated to the t_{tweet} version.

To further analyze this phenomena, and to differentiate the steady state from the changing one, we fitted with blue intention strength points in the graphs with the closest linear regression line (red line) to measure its progression through its slope, as shown as well in Figure 70. Evidently, if the slope was negative, this indicates the intention has changed from *current* to *past*. We use both the slope of the fitted regression line and the fitting error to cluster the plots into three different categories: Steady, Changing, and Unknown. The *Steady Intentional behavior* means the slope is small and the fitting error is small, this indicates a resource where the intention did not change across time. The *Changing Intentional behavior* means the slope is negative, indicating a change in intention from current to past across time, with a moderate fitting error. Finally the *Unknown Intentional behavior* is where the regression line fitting error is too high or the 4th class of TIRM, where the resource is not relevant and did not change.
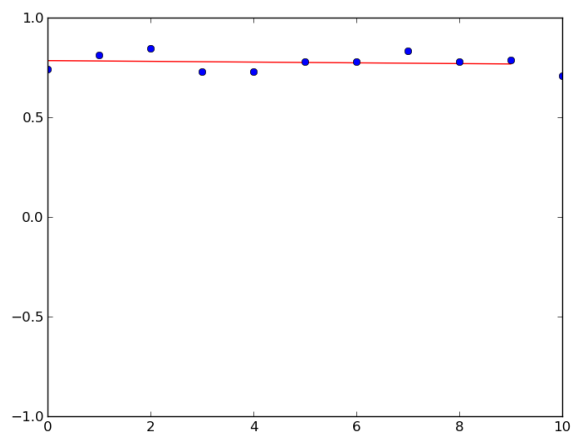
(a) Steady *Current* intention: ≈ 0 slope
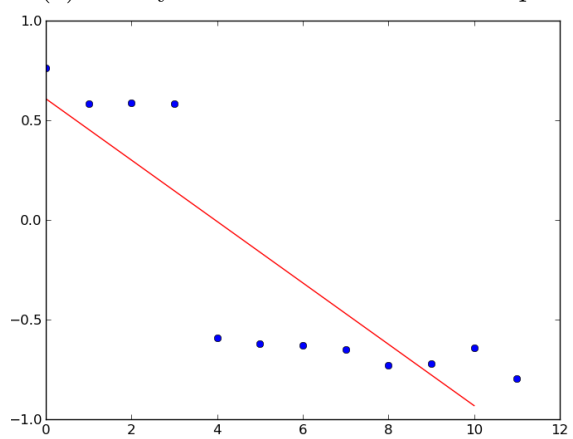
(b) Changing intention: *Current* to *Past*, -ve slope
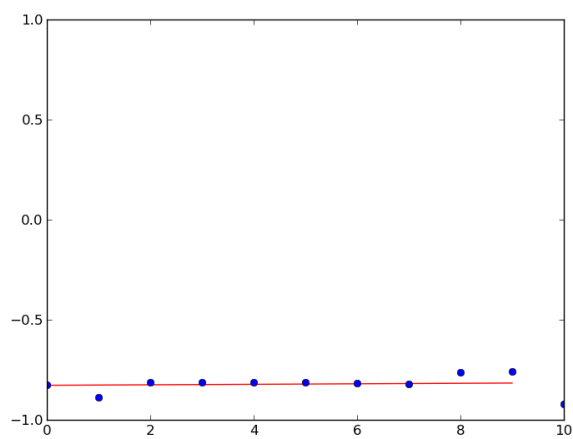
(c) Steady *Past* intention: ≈ 0 slope

Figure 70. The resources' intention strength across time for different behavior categories

Given the slope, intercept, and fitting error, along with the other features, we were able to successfully train a regression classifier to automatically categorize the behavior of a resource across time into either one of these three categories. We performed a 10-fold cross-validation, and the classifier correctly classified 89% of the dataset as shown in Table 31. We were able to identify the behavioral class of intention given the knowledge of the state of the resource and the social network around it through time; the next step was to validate the viability of identifying these classes given only the information available at t_{tweet}.

With our model from the previous stage, we filtered out all the longitudinal temporal features and kept only the features extracted from the tweet and the current version of the resource at t_{tweet}. We retrained the classifier using these limited features and it correctly classified 77% of the dataset. Although, this percentage is lower than the prior percentage of 89% with the full knowledge of the resource in time as expected, it still indicates the viability of predicting the temporal intention progression, given only the knowledge of the tweet at posting time and the state of the resource at t_{tweet}, as shown in Tables 31 and 32.

	10-Fold Cross-Validation Testing	
Model	**With all Features**	**With the tweet and the resource at t_{tweet}**
Mean Absolute Error	0.15	0.22
Relative Absolute Error	34.11	50.57
Kappa Statistic	0.84	0.65
Incorrectly Classified	10.94%	23.32%
Correctly Classified	89.06%	76.68%

Table 31. Results of 10-fold cross-validation for predicting intention behavior strength across time

In other words, given only the information about the resource and the tweet available at the time of authoring a tweet, we can predict for the author's temporal intention and its likelihood of change with 77% accuracy.

Returning to our tweet examples and as shown in Table 33, in the tweet in Figure 68a, the model predicted a change in intention from current to past with

	Precision	Recall	F-measure
Steady Intention	0.680	0.715	0.697
Changing Intention (Current to Past)	0.912	0.897	0.904
Undefined Intention	0.713	0.688	0.700
Weighted Avg.	0.768	0.767	0.767

Table 32. Intention behavior prediction classifier

60% probability. While in the tweet in Figure 68b, the model predicted a 60% probability of steady-current intention. Furthermore, for the third tweet, our model predicted a steady behavior with a 50% probability.

Example Tweet	**Classified Behavior**	**Probability**
check out our latest news at http://bit.ly/1xC7MhK #PolyU	Steady-Current	60%
@heathermeeker The media just lost interest, the WHO has been releasing regular flu A(H1N1) updates, latest is #47 http://bit.ly/whodu	Changing	60%
The Real Secret to Becoming a Popular Blogger http://bit.ly/16OY7q via @FreelanceSw	Steady-Past	50%

Table 33. Tweet examples of the behavior classes

This prediction will give the author sufficient information to choose to just post the tweet or take a snapshot of the resource and push it into one of the public archives and link to that snapshot instead of the assigned URI to maintain the consistency. This prediction will have implication on maintaining the consistency of the conveyed information on the web and will help enrich the archived content of a multitude of resources by crowdsourcing the preservation task.

8.3 SUMMARY

With the quantified intention measure, we analyzed the progress of intention of a (tweet-resource) pair across time from t_{click} back to when it was tweeted at t_{tweet}. This analysis is utilized in the prediction process of the intention at t_{tweet} by gauging

the intention behavior as a function of time.

We started by analyzing the progression of intention through time. We analyzed the progress of intention of a (tweet, resource) pair across time from t_{click} back to when it was tweeted at t_{tweet}. Using our SNAP dataset, we simulated the intention analysis over the period of 3.5 years from June 2009 to January 2013 to observe the intention strength change across time. We observe three different classes of behavior:

- Stable intention (i.e., does not change across time)

- Changing intention (i.e., intention was for the current version then changed to the past version through time)

- Undefined intention (i.e., the information extracted does not provide enough evidence)

We used these observations to fit regression lines to calculate the slopes and intercepts of intention to detect the progression scheme through time. With this knowledge of intention behavior, we trained a classifier to identify these three classes across time. Furthermore, we eliminated all the features acquired in later stages after the posting time and kept only the information available at time t_{tweet}. Given the hypothesis that people's intentions in posting social content determine their writing styles, and such intentions can be characterized by the content and linguistic features of tweets [226], we argue that given these uncovered linguistic features along with the features mined from the resource's current state we can predict the temporal intention behavior of a tweet-resource pair at the initial tweet time with good accuracy. We utilized these features in our previous dataset across time and modeled the change of intention with a success of 89%. Finally we predicted this change or steadiness of intention at t_{tweet} by using only the features that are readily available at t_{tweet} from both the tweet and the resource and were able to successfully predict this intention with 77% accuracy. Giving the authors enough information to aid them to either re-write the tweet with the knowledge of change or push a snapshot of the resource to one of the public archives and link to it instead will help in maintaining the temporal consistency and enriching the archives at the same time.

CHAPTER 9

USING INTENTION IN THE ACTIVE PRESERVATION OF THE SOCIAL WEB

> "I am enough of an artist to draw freely upon my imagi-
> nation. Imagination is more important than knowledge.
> Knowledge is limited. Imagination encircles the world."
> — Albert Einstein

Intention is a fluid, subjective, and ever-changing notion, but at the culmination of this research we were able to identify the temporal intention of social posts at various points in time. Beyond identification, and utilizing an arsenal of features extracted from various archival, social, exsitential, and structural facets of the relationship between the post and the resource, we were able to model this intention and train a classifier to mimic the human perception of intention through time. The model was trained utilizing assignments by subjects in the form of HITs on Amazon's Mechanical Turk. Furthermore, we successfully derived a formula to quantify intention strength and by adding the time dimension to the modeling of intention we were able to perform predictions with high accuracy. Given a tweet with a Bitly shortened URI, we were able to predict the intention steadiness and change of the resource at the time of the tweet authoring along with the prediction confidence.

In this chapter, and utilizing our trained prediction model, we propose a framework of tools to maintain the temporal consistency of shared content for readers and authors. As a summation to this research we focus on three targets:

1. Providing a proof of concept tools for authors and readers to ensure the temporal consistency of intention at times t_{tweet}, t_{click}.

2. Enhance the intention model with continuous underlying feedback and build a larger, sustainable intention corpus collected for research purposes from anonymized user logs.

3. Enrich the web's archived content by seemless pushes to the public web archives when the intention is predicted to change.

In the following sections we discuss each target and where it fits in the proposed framework.

9.1 PREDICT: TEMPORAL CONSISTENCY THROUGH TOOLS

We demonstrate possible implementations of TIRM for tools for both readers and authors. We start by the first prototype "Hover Archive" where we experiment by providing users with archival and clicklog information during their regular browsing. Then we build on that in our second prototype "Archive Shortner" where we merge archiving with shortening in a small implied step as follows.

9.1.1 READER PROTOTYPE 1: HOVER ARCHIVE

While tools for authors will be helpful, their large-scale adoption is likely far in the future. Regardless, we have an immense corpus of social media that predates this functionality, so we need tools that allow us to infer temporal intention during both interactive or batch replay. The key direction here is to provide the user (either author or reader) with time-based context in regards to the post and its associated resource. This prototype superceeded TIRM and with it we wanted to explore the effectiveness of providing archival and content change contextual information to the user during their usual social browsing with minimal intrusion.

From observation, we noted that social media users do not prefer to use excessive add-on tools, as they tend to be distracting. A good compromise is to start with TipTip which is a javascript plugin developed by Drew Wilson, which will create a custom tooltip to replace the default browser tooltip. The tooltip appears when the user hovers with the mouse more than a second on a certain word or sentence. This tooltip can envelope any desired textual or image-based content. It is extremely lightweight and it detects the edges of the browser window and will make sure the tooltip stays within the current window size. It is completely customizable as well via CSS, so we modified it to trigger when the user hovers with the mouse on a URI in a tweet on their feed for more than one second. To make it cross-browser we encapsulated the tiptip tool and our code into a userscript and tested it in 2011. It was compatible with Firefox (with Grease Monkey), Opera 8+ (with embedded support for userscripts), Chrome (with limited embedded support for userscripts),
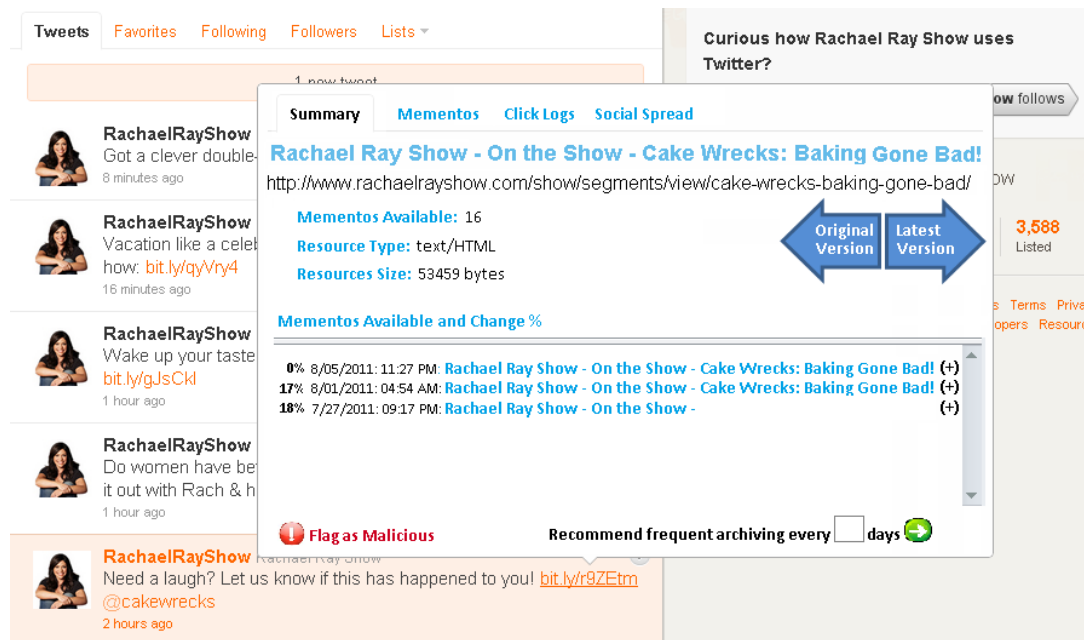
Figure 71. Hovering version of the application displaying the available mementos and resolving the target of the shortned URI

Internet Explorer (using IEPro7, and Grease Monkey for IE), Safari 5+ (with Grease Kit).

Figure 71 shows an interface that queries multiple archives through a Memento Aggregator while reading a Twitter stream. By hovering with the mouse on any shortened URI, it allows the reader to explore extra information in regards to the resource, like the change percentage between the live version and the other mementos of the resource. One of the main purposes of this prototype is to surface archived versions of linked resources in social media and make the users aware of them. The prototype also enables the user to be active by enabling them to opt a resource as archive-worthy and submit it to one of the public archives or flag it as malicious. If it was a Bitly shortened URI the tooltip will show the click logs, total referencing websites and countries, and others.

9.1.2 AUTHOR-READER PROTOTYPE 2: ARCHIVE SHORTNER

The second prototype is targeted to the author. We wanted to analyze the possibility, at the time of posting a tweet, of taking a snapshot of the resource, pushing it into the public archives, creating a memento, bundling the original URI of the resource and the current memento in one package and posting this bundle instead. We developed a server-hosted social archiving service and named it Archive Shortner. As the name entitles, it performs an underlying archiving process while normally shortening the URI via shortners like Bitly as follows:

- Pushes the resource (`www.cs.odu.edu/~mln/`) to a public web archive (we currently use Archive.today, but other services are available) that immediately creates a memento along with a thumbnail (`archive.today/7UODo`).

- Creates a shortened URI for the resource's original URI (`bit.ly/1a31fHg`)

- Creates a second shortened URI for the memento (`bit.ly/1dqSfw1`)

- Creates a third shortened URI (`bit.ly/1cfXc4y`) that points to a service that takes the first and second URIs as arguments (`ws-dl.cs.odu.edu/s?o=bit.ly/1a31fHg&m=bit.ly/1dqSfw1`); this third link is what is sent to Twitter.

Figure 72 shows an iTunes style interface that complements the shortening process described above. Upon clicking on our shortened Bitly URI on Twitter, it performs a redirect to the current page but inserts a header banner cover flow that displays other versions of the resource along with a highlighted version which has the closest memento-datetime to the tweet indicating a past version. The webpages are in the form of successive thumbnails in the display. Alsum and Nelson have performed preliminary investigations about visually summarizing TimeMaps [235]. In this research, they evaluated how HTML can be used to predict changes in thumbnails of mementos, so that k thumbnails can be chosen from a TimeMap of N mementos (where k can be dozens, and N thousands), so changes of a page through time can be summarized without the time and space requirements of generating all thumbnails, not to mention the cognitive load on the user by returning potentially thousands of thumbnails even if they existed.
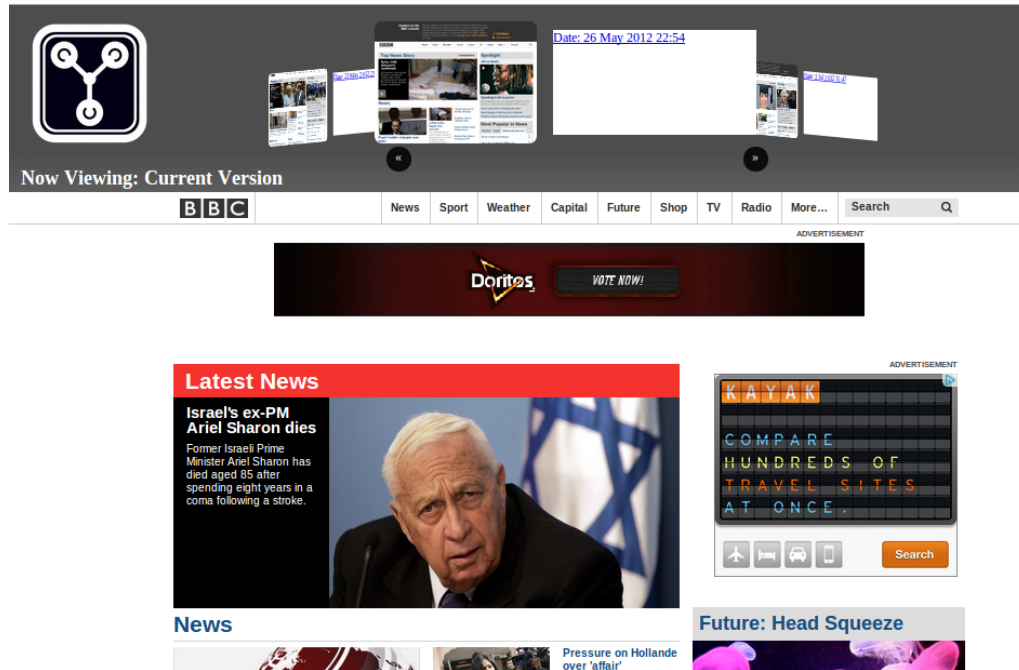
Figure 72. Using our archive shortner and clicking on a link pointing to current BBC front page displayed below, top banner shows thumbnails to archived snapshots, center thumbnail pointing to closest thumbnail to t_{tweet}

9.2 FRAMEWORK: TWITTER ORACLE

With the observations from the previous prototypes and the trained model we proceed in developing our framework. The framework will be divided into two stages based on the two targetted user actors: the author and the reader.

9.2.1 INTENTION ORACLE API

We built a proof-of-concept class prediction service which implements the prediction model described in Section 8.2. The service takes a tweet with a URI shortened via Bit.ly and extracts the necessary features after downloading content and then predicts the behavioral class of the tweet. For the time being, it classifies if the resource is more likely to be in a steady state of intention or a changing state of intention. The service interface is shown in Figure 73 and a sample JSON-encoded response obtained in correspondence to the three tweet examples are demonstrated in Table 33 and Figure 74, respectively.
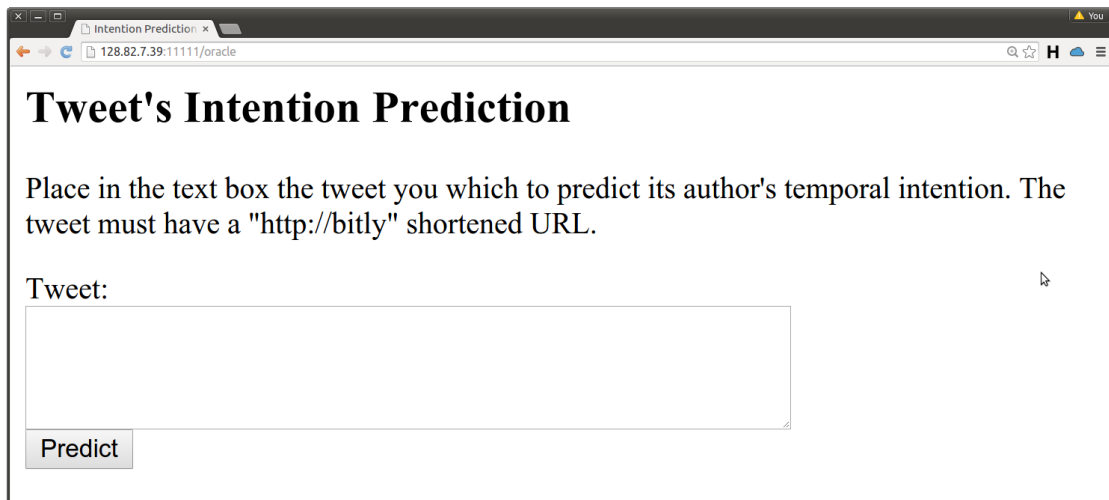
Figure 73. Intention Oracle API service

9.2.2 STAGE 1: THE AUTHOR

At this stage, the browser plugin triggers when the user is writing a new tweet. It executes in the background,and extracts the embedded URI and the related resource properties. It also builds the feature vector for the model in stages based on when the features are calculated. The model generates the prediction and the percentage of confidence. Then when the user presses the tweet button the module presents the options to the author after notifying them with the intention temporal prediction and the corresponding confidence level. The author has the option of:

- Taking a snapshot of the webpage and post a link to the copy.

- Send the URI as it is.

- Bundle the snapshot and the current version and let the reader decide.

The bundle as we will see in stage 2 is very similar to the Archive Shortener discussed in Section 9.1.2.

After executing the author's choice, the anonymized actions are logged, along with the feature vectors. This acts as if it was an assignment from Mechanical Turk like the ones we used to train the model. So in other terms, the model undergoes continuous retraining through feedback. The key is to provide the information in

```
{
  "Tweet Analyzed": "Check out our latest news
  at http://bit.ly/1xC7MhK #PolyU",
  "Bitly Extracted": "http://bit.ly/1xC7MhK",
  "Original Resource URL": "http://www.fb.
  polyu.edu.hk/content/10505/index.html",
  "State": "Steady, Not changing",
  "Prediction": "Predicted Steady intention
  for the resource with 60.0% confidence",
  "Confidence": "60.0"
}

{
  "Tweet Analyzed": "@heathermeeker The media
  just lost interest, the WHO has been
  releasing regular lu A(H1N1) updates,
  latest is #47 http://bit.ly/whodu",
  "Bitly Extracted": "http://bit.ly/whodu",
  "Original Resource URL": "http://www.who.
  int/csr/don/en/",
  "State": "Unsteady, Changing",
  "Prediction": "Predicted Unsteady intention
  observed for the resource, recommend
  preservation with 60.0% confidence",
  "Confidence": "60.0"
}

{
  "Tweet Analyzed": "The Real Secret to
  Becoming a Popular Blogger http://bit.ly/
  16OY7q via @FreelanceSw",
  "Bitly Extracted": "http://bit.ly/16OY7q",
  "Original Resource URL": "http://www.
  copyblogger.com/popular-blogger/",
  "State": "Steady, Not changing",
  "Prediction": "Predicted Steady intention
  for the resource with 50.0% confidence",
  "Confidence": "50.0"
}
```

Figure 74. JSON objects resulting from the Intention Oracle API

a seamless way to the author and require only one click to post to overcome the cognitive load on them. Figure 75 shows a prototype of what the author will see when they hit post tweet.
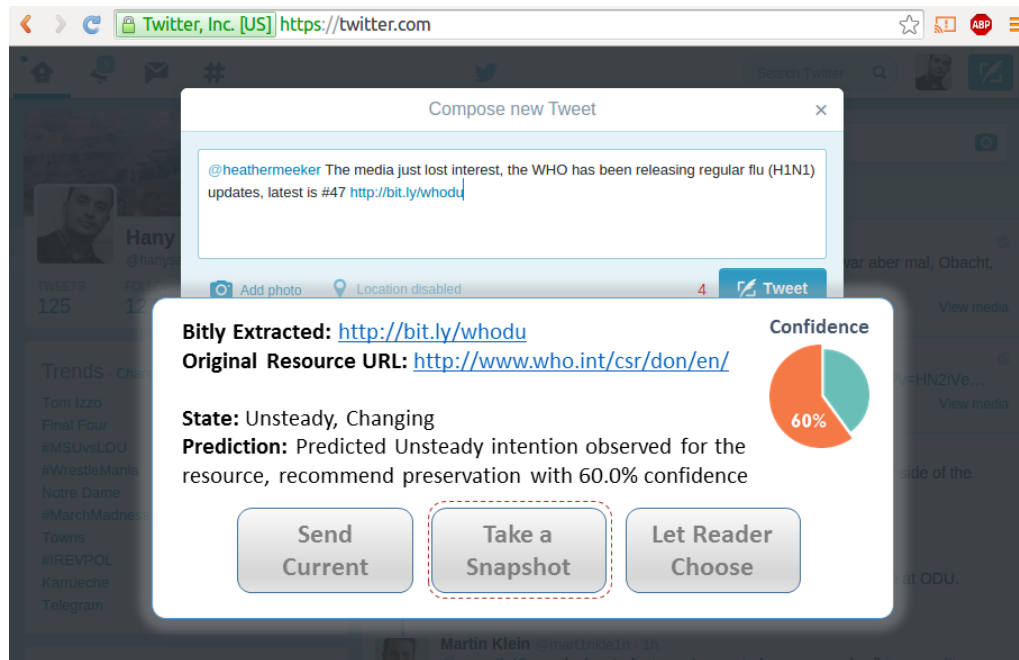
Figure 75. Twitter Oracle Framework: Author-side module

The calculated prediction and confidence are shown to the author along with the recommended course of action. The author is presented with three choices for the intention they wish to convey, each in the form of a button with the labels: send current version, take a snapshot, or let the reader choose.

If the author picked the "current version" option, the module will just shorten the URI through Bitly and post the tweet with that shortened URI. If the author picked the "send a snapshot" option then the module will first push a snapshot into Archive.today. The resulting URI to the snapshot is then in turn shortened via Bitly and posted in the tweet. Finally, if the author decided to "let the reader decide" the module will push a snapshot to the archive, shorten the resultant URI, shorten the original URI, bundle them together and send them to our server. The server generates the thumbnails and the surrounding memento thumbnails, and creates the header banner with the cover flow view of the thumbnails along with the resultant

prediction from our model and the resultant recommendation for which version would better convey the author's intention. This header banner will be injected in the HTML page. This will provide the reader with enough temporal context to maintain the consistency of what the author wanted to convey. We also log all the clicks and navigation the reader does in the header banner along with the final version that the reader decided to read. We also give the reader tools to opt for the resource to be archived, flag as incorrect, or flag as malicious.

9.3 SUMMARY

We demonstrate possible paths for developing tools that will implement TIRM and preserve temporal intention. We demonstrated a prototype API service implementation of our intention prediction model to be utilized on tweets with links shortened via Bitly service. Both the author and the reader will have the knowledge, tools, and the ability to define the intention, get temporal prediction and the corresponding recommendation to maintain the temporal consistency of the social story. They can choose between the version at t_{tweet} and t_{click}, get recommendations, and provide feedback. The feedback collected in the form of usage logs along with the anonymized feature vectors extracted from the tweet-resource pair will be utilized in continous retraining of the model to enhance it. Furthermore, this intention-based collected dataset will be published in the public domain to be utilized by researchers in the field of study. Since there is no other dataset of human behavioral intention in regards to time in the scope of social media or related venues, this dataset is a significant contribution of this dissertation. Finally, by integrating the resource on demand archival in our framework we will enhance the collective quality and quantity of archived content in the public archives. By distributing the archival task to users who navigate the social web we increase the quantity of archived content by simply taking snapshots while we increase the quality of archived content as users will take snapshots of resources that they witness or suspect to change. Resources of high quality will get archived more often than spam pages, and furthermore, users will inject diversity to the archived content by adding social content. Nowadays, the majority of archived content is collected by web crawlers, which in some cases take a while to keep up with the fast changing social content. This increase in the quality of the archive is a desired side effect of socializing the archival process and covering the third and final of our proposed targets.

CHAPTER 10

CONCLUSIONS AND FUTURE WORK

"Not every end is the goal. The end of a melody is not
its goal: but nonetheless, had the melody not reached
its end it would not have reached its goal either. A
parable." — Friedrich Nietzsche

10.1 CONCLUSIONS

Everyday, millions of users author and share content on the social web and an-
notate it with textual content (tweets, facebook statuses) and signs of approval or
disapproval (likes, favorites, thumbs up/down, digs). In several cases users redis-
tribute the content into their social circles (retweet, share). As the web is ever-
changing, in several occasions we have proved that this content, which was shared
and reshared on the social web, does not survive the test of time. The content could
be rendered missing, either by deliberate deletion or by accidental server failure or
hosting service closure. More dangerously instead of loss, the content could have
changed through time. This is critical in the scenario where the author posts a
tweet and link to a resource in it and that resource changes after a period of time.
The readers of that social post will not be able to experience what the author has
originally intended to convey in the post. This leads to a problem of temporal
inconsistency in the shared content on the web. Social media is currently consid-
ered the first draft of history, and posts from individuals during historic events,
riots, revolutions, protests and others are of crucial importance as they closely and
collectively narrate those events. The temporal consistency of these posts is impor-
tant for historic replaying of the events and to demonstrate how they have evolved
through time. For example, posts about the Egyptian revolution of 2011 have been
collected to narrate the protests. This has similarly been the case in the London ri-
ots, Syrian uprising, Tunisian revolution, Occupy Wall Street movement, and others
in the last couple of years. Several researchers conducted experiments on historic

event-related tweet collections to have a better understanding of the information diffusion in this context like the work of Starbird et al. on collections from the Arab Spring and natural disasters [236, 31, 30, 34, 29]. As a side project to this dissertation we worked with Alex Hanna, from the Sociology department at the University of Wisconsin-Madison, on analyzing the tweet datasets related to April 6th Youth Movement during the Egyptian Revolution of 2011 [237]. These studies prove the need of maintaining the temporal consistency in social posts collections for sociologists, historians, scientists, and others.

This dissertation presents the problem of the temporal intention inconsistency of shared content on the web and its effects. In it, we started by quantifying the amount lost and changed on the web and we were able to derive a prediction of the lost content as a function of time. Since we are dealing with the concept of time on the web, we analyzed the archived web content and calculated estimates of how much of the web is archived from different sources. We also proposed a method to find viable replacements from the live web for the missing resources based on their social annotations represented in the form of tweets. As for the changing resources we performed a longitudinal study to regularly gauge the change in these resources from the date of posting through time, and determining when this change does occur and at what rate. Since we needed to know how long a web resource has been on the web prior to its archival, change, or loss, we devised a method to "carbon date" or estimate the age of the resource on the web derived from its social, functional, and archival existence.

Next we analyzed the user aspect of the problem. Temporal intention proved to be a non-trivial concept to gauge after we conducted several experiments on subjects using Amazon's Mechanical Turk. With the goal of detecting this temporal intention, we devised a mapping model to convert the temporal intention problem into two simpler problems of relevancy and change. We called this model TIRM (or the Temporal Intention Relevancy Model). After successfully proving its viability to measure intention, we built a dataset of tweet-resource pairs annotated with the corresponding relevance, change, and in turn temporal intention derived with the aid of five different turkers per instance. With this dataset, we extracted 65 different features and utilized them to build a machine learning model to classify intention. We validated the accuracy and viability of this model by testing it on several extended datasets. We needed a measure to quantify intention so we devised

a corresponding formula to measure intention as a normalized value from -1.0 (denoting past intention) to 1.0 (denoting current intention). Furthermore, we utilized the model and the intention strength calculation along side archived snapshots of the resources spanning the period of 3.5 years to calculate the intention strength through time. With these calculations we consequently plotted the features and the corresponding calculated intention strength through time, and we utilized these calculations as features to train a model to define to which class of steadiness or change this intention belongs. With this trained prediction classifier we were able to predict the class of intention change or steadiness all the way back to the original time of posting the tweet t_{tweet}. This prediction would be extremely useful to provide the authors with knowledge of the state of the tweet-resource pairs they are posting through time. With this knowledge the author could opt for the resource to be archived and link to the archived version instead. Furthermore in the proposed framework, even with no interaction from the author, the model would automatically push the current state of the resource to the public archives when it predicts an inconsistency in intention.

This leads to maintaining the temporal consistency of the shared content and thus helps to save the first draft of history. Furthermore and as a side effect, this model will help in distributing the task of choosing what to archive to users instead of just institutions like the Internet Archive. This distribution will enhance the archived content of the web in both quantity and quality: in quantity by archiving more resources, and in quality by choosing shared content of certain social importance (hence shared), and opting for regularly changed content to be consistently and properly archived and linked. As a proof of concept we proposed a framework of tools that would be built in the browsers and provide this seamless, enriched, and consistent experience in authoring and reading of shared content.

10.2 CONTRIBUTIONS

This dissertation contributes to the fields of web archival, social media analysis, user-behavioral studies, and content analysis as follows:

1. We quantified how much of the web is archived and where it is archived from various sources and datasets and ranges from 17% to 90% accordingly.

2. We quantified how much of the web is missing by deletion or loss. We obtained a collection of tweet-resource pairs about the Egyptian Revolution of 2011 and we found 11% to be missing after just one year. Moreover, we collected a dataset of five other events spanning 3 years and added them to the Egyptian Revolution dataset and we unravelled a relationship between the amount missing of the web and time to be around 11% in the first year and about 7% loss every following year, thus enabling us to predict this loss in the future. We also published this dataset we used in calculation.

3. We analyzed the phenomenon of archived content loss and deduced an average of 8% percent per year will disappear from the archives. Also we analyzed another phenomenon of reappearance on the live web and we calculated an average of 6.5% of the resources would reappear after deemed missing.

4. We devised a reliable method of estimating the creation date of web resources (carbon date) which successfully estimated the correct creation dates for 76% of the test sets and we published these test sets openly to be used by the scientific community in validating creation dates.

5. We analyzed how social content change through time from the date of its first posting on the web by conducting a longitudinal study on a dataset of freshly extracted tweet-resource pairs and capturing the content hourly for an extended period of time. After just one hour, $\sim$4% of the resources have changed by $\geq$30% while after a day the change rate slowed to be $\sim$12% of the resources changed by $\geq$40%.

6. We conducted user-behavioral analysis experiments and built a dataset of 1,124 instances to detect temporal intention in social media and made it publicly available to the research community.

7. We proposed our Temporal Intention Relevancy Model (TIRM) where we transform temporal intention in to simpler subproblems of relevancy and change.

8. We successfully modeled human temporal intention using our proposed model TIRM and evaluated it to yield a 90.27% success rate.

9. We extended TIRM and used it to build a time based model to predict temporal intention change or steadiness at the time of posting t_{tweet} with 77%

accuracy.

10. We built a service API around this model to provide predictions along with confidence measures to the public.

11. We proposed a framework of tools that would work seamlessly in the users' browsers by utilizing our model to maintain the temporal consistency of shared content, provided educated predictions of change to the authors, provided intention recommendations to the readers, and distributed the task of archival selection.

12. We proposed the utilization of this model to continuously retrain TIRM and build an anonymized large temporal intention dataset based on the anonymized user logs to be used openly for research purposes in the scientific community.

10.3 FUTURE WORK

This dissertation identifies the problem of detecting, modeling, and predicting human temporal intention in social media and paves the way to the data acquisition and analysis of this phenomenon. However, this work is far from done and in this section we propose the various angles where we will proceed to explore the problem and proposed solutions.

First, the collection of the snapshots of the resources in the longitudinal study is underway at the time of writing this dissertation. We propose to continue this collection for the next year, which will give us a complete insight of how a resource would change in an extended period of time and at which rate. This dataset will also be publicly distributed and would be useful in rate of change analysis, social spread analysis, clicklog and access analysis, and crawling refresh policies analysis as well.

Second, with the large-scale dataset collected from the framework tools, we need to extend TIRM and enhance it, especially in the prediction angle, which is still rudimentary.

Third, we need to develop the proposed framework, test it, publish it, and sustain it for the public usage. A separate user-experience study should be conducted to gauge usability, gamefication possibilities, and other aspects to encourage the regular user to use it effectively on a daily basis.

REFERENCES

[1] Herbert Van de Sompel. Memento: Updated Technical Details. `http://www.slideshare.net/hvdsomp/memento-updated-technical-details-february-2010`, March 2010.

[2] Malcolm Otter and H. Johnson. Lost In Hyperspace: Metrics and Mental Models. *Interacting With Computers*, 13(1):1–40, 2000.

[3] Facebook.Com. Facebook Official Fact Sheet. `http://newsroom.fb.com/content/default.aspx?NewsAreaId=22`, 2012.

[4] Twitter. Twitter Usage and Company Facts. `https://about.twitter.com/company`, 2014.

[5] N. Megiddo and K.S. Mccurley. Efficient Retrieval Of Uniform Resource Locators, October 18 2005. US Patent 6,957,224.

[6] Frank McCown and Michael L. Nelson. Agreeing To Disagree: Search Engines and Their Public Interfaces. In *Proceedings Of The 7th ACM/IEEE-CS Joint Conference On Digital Libraries*, JCDL '07, pages 309–318, New York, NY, USA, 2007. ACM.

[7] Herbert Van de Sompel, Michael L. Nelson, Robert Sanderson, Lyudmila Balakireva, Scott Ainsworth, and Harihar Shankar. Memento: Time Travel For The Web. Technical Report arXiv:0911.1112, 2009.

[8] Koen Holtman and Andrew Mutz. Transparent Content Negotiation In HTTP, Internet RFC-2295, 1998.

[9] Tim Berners-Lee. Web Architecture: Generic Resources. `http://www.w3.org/DesignIssues/Generic.html`, 1996.

[10] Michael L. Nelson Herbert Van de Sompel and Robert Sanderson. HTTP Framework For Time-Based Access To Resource States – Memento. Internet RFC-7089, 2013.

[11] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What Is Twitter, A Social Network Or A News Media? In *Proceedings Of The 19th*

International Conference On World Wide Web, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[12] M. E. J. Newman and Juyong Park. Why Social Networks Are Different From Other Types Of Networks. *Physical Review E*, 68(3):036122+, September 2003.

[13] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why We Twitter: Understanding Microblogging Usage and Communities. In *Proceedings Of The 9th WebKDD and 1st Sna-Kdd 2007 Workshop On Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM.

[14] Dejin Zhao and Mary B Rosson. How and Why People Twitter: The Role That Micro-Blogging Plays In Informal Communication At Work. In *Proceedings Of The ACM 2009 International Conference On Supporting Group Work*, GROUP '09, pages 243–252, New York, NY, USA, 2009. ACM.

[15] Meeyoung Cha, Alan Mislove, and Krishna P. Gummadi. A Measurement-Driven Analysis Of Information Propagation In The Flickr Social Network. In *Proceedings Of The 18th International Conference On World Wide Web*, WWW '09, pages 721–730, New York, NY, USA, 2009. ACM.

[16] J Yang and S Counts. Predicting The Speed, Scale, and Range Of Information Diffusion In Twitter. In *4th International AAAI Conference On Weblogs and Social Media (ICWSM)*, May 2010.

[17] Mario Cataldi, Luigi Di Caro, and Claudio Schifanella. Emerging Topic Detection On Twitter Based On Temporal and Social Terms Evaluation. In *Proceedings Of The Tenth International Workshop On Multimedia Data Mining*, MDMKDD '10, pages 4:1—-4:10, New York, NY, USA, 2010. ACM.

[18] Yan Chen, Hadi Amiri, Zhoujun Li, and Tat-Seng Chua. Emerging Topic Detection For Organizations From Microblogs. In *Proceedings Of The 36th International ACM SIGIR Conference On Research and Development In Information Retrieval*, SIGIR '13, pages 43–52, New York, NY, USA, 2013. ACM.

[19] Jianshu Weng and Bu-Sung Lee. Event Detection In Twitter. In *5th International AAAI Conference On Weblogs and Social Media (ICWSM)*, May 2011.

[20] Michael Mathioudakis and Nick Koudas. Twittermonitor: Trend Detection Over The Twitter Stream. In *Proceedings Of The 2010 ACM SIGMOD International Conference On Management Of Data*, pages 1155–1158. ACM, 2010.

[21] James Benhardus and Jugal Kalita. Streaming Trend Detection In Twitter. *International Journal Of Web Based Communities*, 9(1):122–139, 2013.

[22] Swit Phuvipadawat and Tsuyoshi Murata. Breaking News Detection and Tracking In Twitter. In *Proceedings Of The 2010 IEEE/WIC/ACM International Conference On Web Intelligence and Intelligent Agent Technology - Volume 03*, WI-IAT '10, pages 120–123, Washington, DC, USA, 2010. IEEE Computer Society.

[23] Wei Xie, Feida Zhu, Jing Jiang, Ee-Peng Lim, and Ke Wang. Topicsketch: Real-Time Bursty Topic Detection From Twitter. In *Data Mining (ICDM), 2013 IEEE 13th International Conference On*, pages 837–846, Dec 2013.

[24] Louis Lei Yu, Sitaram Asur, and Bernardo A. Huberman. What Trends In Chinese Social Media. Technical Report arXiv:1107.3522, 2011.

[25] Xun Zhao, Feida Zhu, Weining Qian, and Aoying Zhou. Impact Of Multimedia In Sina Weibo: Popularity and Life Span. In Juanzi Li, Guilin Qi, Dongyan Zhao, Wolfgang Nejdl, and Hai-Tao Zheng, editors, *Semantic Web and Web Science*, Springer Proceedings in Complexity, pages 55–65. Springer New York, 2013.

[26] Tao Chen, Dongyuan Lu, Min-Yen Kan, and Peng Cui. Understanding and Classifying Image Tweets. In *Proceedings Of The 21st ACM International Conference On Multimedia*, MM '13, pages 781–784, New York, NY, USA, 2013. ACM.

[27] Malcolm Gladwell. Small Change: Why The Revolution Will Not Be Tweeted. `http://www.newyorker.com/magazine/2010/10/04/small-change-3`, 2010.

[28] Leo Mirani. Sorry, Malcolm Gladwell, The Revolution May Well Be Tweeted. `http://www.theguardian.com/commentisfree/cifamerica/2010/oct/02/malcolm-gladwell-social-networking-kashmir`, 2010.

[29] Kate Starbird and Leysia Palen. (How) Will The Revolution Be Retweeted?: Information Diffusion and The 2011 Egyptian Uprising. In *Proceedings Of The*

ACM 2012 Conference On Computer Supported Cooperative Work, CSCW '12, pages 7–16, New York, NY, USA, 2012. ACM.

[30] Kate Starbird, Leysia Palen, Amanda L Hughes, and Sarah Vieweg. Chatter On The Red: What Hazards Threat Reveals About The Social Life Of Microblogged Information. In *Proceedings Of The 2010 ACM Conference On Computer Supported Cooperative Work*, CSCW '10, pages 241–250, New York, NY, USA, 2010. ACM.

[31] Sarah Vieweg, Amanda L. Hughes, Kate Starbird, and Leysia Palen. Microblogging During Two Natural Hazards Events: What Twitter May Contribute To Situational Awareness. In *Proceedings Of The 28th International Conference On Human Factors In Computing Systems - CHI '10*, page 1079, New York, New York, USA, April 2010. ACM Press.

[32] Yan Qu, Chen Huang, Pengyi Zhang, and Jun Zhang. Microblogging After A Major Disaster In China: A Case Study Of The 2010 Yushu Earthquake. In *Proceedings Of The ACM 2011 Conference On Computer Supported Cooperative Work*, CSCW '11, pages 25–34, New York, NY, USA, 2011. ACM.

[33] Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. Safety Information Mining - What Can NLP Do In A Disaster. In *In Proceedings Of The 5th International Joint Conference On Natural Language Processing (IJCNLP'11)*, pages 965–973, 2011.

[34] Kate Starbird and Leysia Palen. "Voluntweeters": Self-Organizing By Digital Volunteers In Times Of Crisis. In *Proceedings Of The 2011 Annual Conference On Human Factors In Computing Systems*, CHI '11, pages 1071–1080, New York, NY, USA, 2011. ACM.

[35] Alex Burns and Ben Eltham. Twitter Free Iran: An Evaluation Of Twitter's Role In Public Diplomacy and Information Operations In Iran's 2009 Election Crisis. In *Communications Policy and Research Forum 2009*, pages 322–334, November 2009.

[36] Grace Muzny Kate Starbird and Leysia Palen. Learning From The Crowd: Collaborative Filtering Techniques For Identifying On-The-Ground Twitterers

During Mass Disruptions. In *Proceedings Of The 9th International ISCRAM Conference*, ISCRAM '12, Vancouver, Canada, 2012.

[37] Janette Lehmann, Carlos Castillo, Mounia Lalmas, and Ethan Zuckerman. Finding News Curators In Twitter. In *Proceedings Of The 22nd International Conference On World Wide Web Companion*, WWW '13 Companion, pages 863–870, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[38] Janette Lehmann, Carlos Castillo, Mounia Lalmas, and Ethan Zuckerman. Transient News Crowds In Social Media. In *7th International AAAI Conference On Weblogs and Social Media (AAAI-ICWSM)*, 2013.

[39] Gloria Mark, Mossaab Bagdouri, Leysia Palen, James Martin, Ban Al-Ani, and Kenneth anderson. Blogs As A Collective War Diary. In *Proceedings Of The ACM 2012 Conference On Computer Supported Cooperative Work - CSCW '12*, page 37, New York, New York, USA, February 2012. ACM Press.

[40] Alastair J. Gill, Scott Nowson, and Jon Oberlander. What Are They Blogging About? Personality, Topic and Motivation In Blogs. In *3rd International AAAI Conference On Weblogs and Social Media (AAAI-ICWSM)*, May 2009.

[41] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing Web Search Using Social Annotations. In *Proceedings Of The 16th International Conference On World Wide Web*, WWW '07, pages 501–510, 2007.

[42] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Can Social Bookmarking Enhance Search in the Web? In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 107–116, New York, NY, USA, 2007. ACM.

[43] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can Social Bookmarking Improve Web Search? In *Proceedings Of The International Conference On Web Search and Web Data Mining*, WSDM '08, pages 195–206, New York, NY, USA, 2008. ACM.

[44] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. Social Tag Prediction. In *SIGIR '08: Proceedings Of The 31st Annual International ACM SIGIR Conference On Research and Development In Information Retrieval*, pages 531–538, New York, NY, USA, 2008. ACM.

[45] Paul Heymann, Andreas Paepcke, and Hector Garcia-Molina. Tagging Human Knowledge. In *Proceedings Of The Third ACM International Conference On Web Search and Data Mining*, WSDM '10, pages 51–60, New York, NY, USA, 2010. ACM.

[46] Paul Heymann and Hector Garcia-Molina. Contrasting Controlled Vocabulary and Tagging: Do Experts Choose The Right Names To Label The Wrong Things? In *Proceedings Of The Second ACM International Conference On Web Search and Data Mining (WSDM 2009), Late Breaking Results Session*, pages 1–4, February 2009.

[47] Paul Heymann. Final Project: On The Use and Abuse Of Collaborative Tagging Data. 2006. `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.143.619`.

[48] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can All Tags Be Used For Search? In *Proceedings Of The 17th ACM Conference On Information and Knowledge Management*, CIKM '08, pages 193–202, New York, NY, USA, 2008. ACM.

[49] Martin Klein and Michael L. Nelson. Find, new, copy, web, page - tagging for the (re-)discovery of web pages. In *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries*, TPDL'11, pages 27–39, Berlin, Heidelberg, 2011. Springer-Verlag.

[50] Patrick Pantel, Michael Gamon, Omar Alonso, and Kevin Haas. Social Annotations: Utility and Prediction Modeling. In *Proceedings Of The 35th International ACM SIGIR Conference On Research and Development In Information Retrieval*, SIGIR '12, pages 285–294, New York, NY, USA, 2012. ACM.

[51] Jiyin He, Maarten de Rijke, Merlijn Sevenster, Rob Van Ommering, and Yuechen Qian. Generating Links To Background Knowledge: A Case Study Using Narrative Radiology Reports. In *Proceedings Of The 20th ACM International*

Conference On Information and Knowledge Management, CIKM '11, pages 1867–1876, New York, NY, USA, 2011. ACM.

[52] Rada Mihalcea and andras Csomai. Wikify!: Linking Documents To Encyclopedic Knowledge. In *Proceedings Of The Sixteenth ACM Conference On Conference On Information and Knowledge Management*, CIKM '07, pages 233–242, New York, NY, USA, 2007. ACM.

[53] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. Linden: Linking Named Entities With Knowledge Base Via Semantic Knowledge. In *Proceedings Of The 21st International Conference On World Wide Web*, WWW '12, pages 449–458, New York, NY, USA, 2012. ACM.

[54] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Zencrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques For Large-Scale Entity Linking. In *Proceedings Of The 21st International Conference On World Wide Web*, WWW '12, pages 469–478, New York, NY, USA, 2012. ACM.

[55] Marc Bron, Bouke Huurnink, and Maarten de Rijke. Linking archives using document enrichment and term selection. In *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries*, TPDL'11, pages 360–371. Springer-Verlag, Berlin, Heidelberg, 2011.

[56] Silviu Cucerzan. Large-Scale Named Entity Disambiguation Based On Wikipedia Data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 708–716, 2007.

[57] Xianpei Han and Jun Zhao. Named Entity Disambiguation By Leveraging Wikipedia Semantic Knowledge. In *Proceedings Of The 18th ACM Conference On Information and Knowledge Management*, CIKM '09, pages 215–224, New York, NY, USA, 2009. ACM.

[58] Hakan Ceylan, Ioannis Arapakis, Pinar Donmez, and Mounia Lalmas. Automatically Embedding Newsworthy Links To Articles. In *Proceedings Of The 21st*

ACM International Conference On Information and Knowledge Management - CIKM '12, page 1502, New York, New York, USA, October 2012. ACM Press.

[59] Martin Klein. *Using The Web Infrastructure For Real Time Recovery Of Missing Web Pages*. PhD thesis, Old Dominion University Department of Computer Science, 2011.

[60] Matt Hodkinson. The Best URL Shorteners For Tracking Social Media Success. `http://www.influenceagents.com/matts-chat/url-shorteners-tracking-social-media-success`, Accessed: 2014.

[61] Demetris Antoniades, Iasonas Polakis, Georgios Kontaxis, Elias Athanasopoulos, Sotiris Ioannidis, Evangelos P. Markatos, and Thomas Karagiannis. We.B: The Web Of Short URls. In *Proceedings Of The 20th International Conference On World Wide Web*, WWW '11, pages 715–724, New York, NY, USA, 2011. ACM.

[62] Tim Berners-Lee. Cool Uris Don't Change. `http://www.w3.org/Provider/Style/URI.html`. 1998.

[63] Helen Ashman. Electronic Document Addressing: Dealing With Change. *ACM Computing Surveys*, 32(3):201–212, 2000.

[64] Helen Ashman, Hugh Davis, Jim Whitehead, and Steve Caughey. Missing The 404: Link Integrity On The World Wide Web. In *Proceedings Of The 7th International Conference On World Wide Web - WWW '98*, pages 761–762. ACM Press, 1998.

[65] Hugh C. Davis. Hypertext Link Integrity. *ACM Computer Survey*, 31(4es), December 1999.

[66] Hugh C. Davis. Referential Integrity Of Links In Open Hypermedia Systems. In *Proceedings Of Hypertext '98*, pages 207–216, 1998.

[67] Wallace Koehler. Web Page Change and Persistence — A Four-Year Longitudinal Study. *Journal Of The American Society For Information Science and Technology*, 53(2):162–171, 2002.

[68] Steve Lawrence, David M. Pennock, Gary William Flake, Robert Krovetz, Frans M. Coetzee, Eric Glover, Finn Nielsen, Andries Kruger, and C. Lee

Giles. Persistence Of Web References In Scientific Research. *IEEE Computer*, 34(2):26–31, 2001.

[69] Diomidis Spinellis. The Decay and Failures Of Web References. *Communications Of The ACM*, 46(1):71–77, 2003.

[70] Frank McCown, Sheffan Chan, Michael L. Nelson, and Johan Bollen. The Availability and Persistence Of Web References In D-Lib Magazine. In *5th International Web Archiving Workshop (IWAW'05)*, September 2005.

[71] Michael L. Nelson and B. Danette Allen. Object Persistence and Availability In Digital Libraries. *D-Lib Magazine*, 8(1), 2002.

[72] Kurt Maly, Michael L. Nelson, and Mohammad Zubair. Smart Objects, Dumb Archives: A User-Centric, Layered Digital Library Framework. *D-Lib Magazine*, 5(3), 1999.

[73] E. Russell and J. Kane. The Missing Link: Assessing The Reliability Of Internet Citations In History Journals. *Technology and Culture*, 49(2):420–429, 2008.

[74] Robert Sanderson, Mark Phillips, and Herbert Van de Sompel. Analyzing The Persistence Of Referenced Web Resources With Memento. In *Proceedings Of Open Repositories 2011*, 2011.

[75] Martin Klein, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, and Richard Tobin. Scholarly Context Not Found: One In Five Articles Suffers From Reference Rot. *PLOS One*, 9(12):e115253, 12 2014.

[76] Robert P. Dellavalle, Eric J. Hester, Lauren F. Heilig, Amanda L. Drake, Jeff W. Kuntzman, Marla Graber, and Lisa M. Schilling. Going, Going, Gone: Lost Internet References. *Science*, 302(5646):787–788, October 2003.

[77] John Markwell and David W Brooks. Broken Links: The Ephemeral Nature Of Educational WWW Hyperlinks. *Journal Of Science Education and Technology*, 11:105–108, 2002.

[78] Ziv Bar-Yossef, andrei Z. Broder, Ravi Kumar, and andrew Tomkins. Sic Transit Gloria Telae: Towards An Understanding Of The Web's Decay. In

Proceedings Of The 13th International Conference On World Wide Web, WWW '04, pages 328–337, New York, NY, USA, 2004. ACM.

[79] Norman Paskin. Digital Object Identifiers. *Information Services and Use*, 22(2-3):97–112, 2002.

[80] S. Sun, L. Lannom, and B. Boesch. Handle System Overview. Informational RFC 3650, 2003.

[81] K. Shafer, S. Weibel, E. Jul, and J. Fausey. Persistent Uniform Resource Locators. `http://www.purl.org/`.

[82] William Y. Arms. Uniform Resource Names: Handles, Purls, and Digital Object Identifiers. *Communications of the ACM*, 44(5):68–, May 2001.

[83] Akiyoshi Nakamizo, Toshinari Iida, Atsuyuki Morishima, Shigeo Sugimoto, and Hiroyuki Kitagawa. A Tool To Compute Reliable Web Links and Its Applications. In *Proceedings Of ICDEW '05*, page 1255, 2005.

[84] Atsuyuki Morishima, Akiyoshi Nakamizo, Toshinari Iida, Shigeo Sugimoto, and Hiroyuki Kitagawa. Pagechaser: A Tool For The Automatic Correction Of Broken Web Links. In *Proceedings Of ICDE '08*, pages 1486–1488, 2008.

[85] Errorzilla - Useful Error Pages For Firefox. `http://www.jaybaldwin.com/Blog.aspx?cid=4`.

[86] W. John Wilbur and Karl Sirotkin. The Automatic Identification Of Stop Words. *Journal Of Information Science*, 18(1):45–55, 1992.

[87] Martin F. Porter. An Algorithm For Suffix Stripping. *Program: Electronic Library and Information Systems*, 14(3):130–137, 1980.

[88] Amit Singhal, Chris Buckley, and Mandar Mitra. Pivoted Document Length Normalization. In *Proceedings Of The 19th International ACM SIGIR Conference On Research and Development In Information Retrieval*, SIGIR '96, pages 21–29, 1996.

[89] Karen Spaerck Jones. A Statistical Interpretation Of Term Specificity and Its Application In Retrieval. *Journal Of Documentation*, 28:11–21, 1972.

[90] Stephen E. Robertson. Understanding Inverse Document Frequency: On Theoretical Arguments For IDF. *Journal Of Documentation*, 60(5):503–520, 2004.

[91] S. E. Robertson and K. Sparck Jones. Simple, Proven Approaches To Text Retrieval. Technical Report 356, December 1994, University of Cambridge, Computer Laboratory.

[92] Gerard Salton and Christopher Buckley. Term-Weighting Approaches In Automatic Text Retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[93] Gerard Salton and Chu-Sing Yang. On The Specification Of Term Values In Automatic Indexing. *Journal Of Documentation*, 29:351–372, 1973.

[94] Kazunari Sugiyama, Kenji Hatano, Masatoshi Yoshikawa, and Shunsuke Uemura. Refinement Of TF-IDF Schemes For Web Pages Using Their Hyperlinked Neighboring Pages. In *Hypertext '03: Proceedings Of The Fourteenth ACM Conference On Hypertext and Hypermedia*, pages 198–207, 2003.

[95] Brian D. Davison. Topical Locality In The Web. In *SIGIR '00: Proceedings Of The 23rd Annual International ACM SIGIR Conference On Research and Development In Information Retrieval*, pages 272–279, 2000.

[96] Jeffrey Dean and Monika R. Henzinger. Finding Related Pages In The World Wide Web. *Computer Networks*, 31(11-16):1467–1479, 1999.

[97] Thomas A. Phelps and Robert Wilensky. Robust hyperlinks cost just five words each. Technical Report UCB//CSD-00-1091, University of California at Berkeley, CA, USA, 2000.

[98] Luis Meneses, Himanshu Barthwal, Sanjeev Singh, Richard Furuta, and Frank Shipman. Restoring Semantically Incomplete Document Collections Using Lexical Signatures. In *Proceedings Of The 17th International Conference On Theory and Practice Of Digital Libraries*, TPDL'13, pages 321–332. Springer Berlin Heidelberg, 2013.

[99] Seung-Taek Park, David M. Pennock, C. Lee Giles, and Robert Krovetz. Analysis Of Lexical Signatures For Improving Information Persistence On The World Wide Web. *ACM Transactions On Information Systems*, 22(4):540–572, October 2004.

[100] Xiaojun Wan and Jianwu Yang. Wordrank-Based Lexical Signatures For Finding Lost Or Related Web Pages. In *Proceedings Of The 8th Asia-Pacific Web Conference On Frontiers Of WWW Research and Development*, APWeb'06, pages 843–849, Berlin, Heidelberg, 2006. Springer-Verlag.

[101] Jessica Staddon, Philippe Golle, and Bryce Zimny. Web Based Inference Detection. In *Usenix Security Symposium*, 2007.

[102] Monika Henzinger, Bay-Wei Chang, Brian Milch, and Sergey Brin. Query-Free News Search. In *WWW '03: Proceedings Of The 12th International Conference On World Wide Web*, pages 1–10, 2003.

[103] Martin Klein and Michael L. Nelson. Moved But Not Gone: An Evaluation Of Real-Time Methods For Discovering Replacement Web Pages. *International Journal On Digital Libraries*, 14(1-2):17–38, 2014.

[104] Martin Klein, Jeb Ware, and Michael L. Nelson. Rediscovering Missing Web Pages Using Link Neighborhood Lexical Signatures. In *Proceedings Of The 11th Annual International ACM/IEEE Joint Conference On Digital Libraries*, JCDL '11, pages 137–140, New York, NY, USA, 2011. ACM.

[105] Jeb Ware, Martin Klein, and Michael L. Nelson. An Evaluation of Link Neighborhood Lexical Signatures to Rediscover Missing Web Pages. Technical Report arXiv:1102.0930, CS Department, Old Dominion University, Norfolk, Virginia, USA, 2011.

[106] R. Rivest. The MD5 Message-Digest Algorithm RFC-1321, April 1992.

[107] Donald E. Eastlake 3rd and Paul E. Jones. US Secure Hash Algorithm 1 (SHA1) Internet RFC-3174, 2001.

[108] Moses S. Charikar. Similarity Estimation Techniques From Rounding Algorithms. In *Proceedings Of The Thiry-Fourth Annual ACM Symposium On Theory Of Computing*, STOC '02, pages 380–388, New York, NY, USA, 2002. ACM.

[109] Gurmeet S. Manku, Arvind Jain, and Anish D. Sarma. Detecting Near-Duplicates For Web Crawling. In *Proceedings Of The 16th International Conference On World Wide Web - WWW '07*, pages 141–150. ACM Press, 2007.

[110] Lars Clausen. Concerning ETags and Datestamps. In *4th International Web Archiving Workshop (IWAW04)*, 2004.

[111] Zubin Dalal, Suvendu Dash, Pratik Dave, Luis Francisco-Revilla, Richard Furuta, Unmil Karadkar, and Frank Shipman. Managing Distributed Collections: Evaluating Web Page Changes, Movement, and Replacement. In *Proceedings Of The 4th ACM/IEEE-CS Joint Conference On Digital Libraries*, JCDL '04, pages 160–168, New York, NY, USA, 2004. ACM.

[112] Luis Francisco-Revilla, Frank Shipman, Richard Furuta, Unmil Karadkar, and Avital Arora. Managing Change On The Web. In *Proceedings Of The 1st ACM/IEEE-CS Joint Conference On Digital Libraries*, JCDL '01, pages 67–76, New York, NY, USA, 2001. ACM.

[113] Eytan Adar, Jaime Teevan, Susan T. Dumais, and Jonathan L. Elsas. The Web Changes Everything: Understanding The Dynamics Of Web Content. In *WSDM '09: Proceedings Of The Second ACM International Conference On Web Search and Data Mining*, pages 282–291, 2009.

[114] Junghoo Cho and Hector Garcia-Molina. Effective Page Refresh Policies For Web Crawlers. *ACM Transactions On Database Systems (TODS)*, 28(4):390–426, 2003.

[115] Junghoo Cho and Hector Garcia-Molina. Estimating Frequency Of Change. *ACM Transactions On Internet Technology*, 3(3):256–290, August 2003.

[116] Alexandros Ntoulas, Junghoo Cho, and Christopher Olston. What's New On The Web? The Evolution Of The Web From A Search Engine Perspective. In *WWW '04: Proceedings Of The 13th International Conference On World Wide Web*, pages 1–12, 2004.

[117] Fred Douglis, Anja Feldmann, Balachander Krishnamurthy, and Jeffrey Mogul. Rate Of Change and Other Metrics: A Live Study Of The World Wide Web. In *USITS'97: Proceedings Of The Usenix Symposium On Internet Technologies and Systems On Usenix Symposium On Internet Technologies and Systems*, 1997.

[118] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener. A Large-Scale Study Of The Evolution Of Web Pages. In *WWW '03: Proceedings Of The 12th International Conference On World Wide Web*, pages 669–678, 2003.

[119] K. Radinsky and P. N. Bennett. Predicting Content Change On The Web. *WSDM '13: Proceedings Of The Sixth ACM International Conference On Web Search and Data Mining*, 2013.

[120] M. Ben Saad, Z. Pehlivan, and S. Gançarski. Coherence-Oriented Crawling and Navigation Using Patterns For Web Archives. In *Proceedings Of The 15th International Conference On Theory and Practice Of Digital Libraries: Research and Advanced Technology For Digital Libraries*, TPDL'11, pages 421–433, Berlin, Heidelberg, 2011. Springer-Verlag.

[121] M. Ben Saad and S. Gançarski. Archiving The Web Using Page Changes Patterns: A Case Study. In *Proceedings Of The 11th Annual International ACM/IEEE Joint Conference On Digital Libraries*, pages 113–122, 2011.

[122] M. Spaniol, A. Mazeika, D. Denev, and G. Weikum. "Catch Me If You Can": Visual Analysis Of Coherence Defects In Web Archiving. In *The 9th International Web Archiving Workshop (IWAW 2009) Corfu, Greece, September/October, 2009 Workshop Proceedings*, 2009.

[123] Dimitar Denev, Arturas Mazeika, Marc Spaniol, and Gerhard Weikum. Sharc: Framework For Quality-Conscious Web Archiving. *Proceedings Of The VLDB Endowment*, 2(1):586–597, August 2009.

[124] Venugopalan Ramasubramanian and Emin G Sirer. Perils Of Transitive Trust In The Domain Name System. In *Proceedings Of IMC '05*, pages 35–40, 2005.

[125] Security and Stability Advisory Committee ICANN SSAC. Domain Name Hijacking: Incidents, Threats, Risks, and Remedial Actions. `http://www.icann.org/en/announcements/hijacking-report-12jul05.pdf`, 2005.

[126] Frank McCown, Catherine C. Marshall, and Michael L. Nelson. Why Web Sites Are Lost (and How They're Sometimes Found). *Communications of the ACM*, 52(11):141–145, November 2009.

[127] Frank McCown. *Lazy Preservation: Reconstructing Websites From The Web Infrastructure*. PhD thesis, Old Dominion University Department of Computer Science, 2007.

[128] Martin Klein and Michael L. Nelson. Revisiting Lexical Signatures To (Re-)Discover Web Pages. In *Proceedings Of The 12th European Conference On Research and Advanced Technology For Digital Libraries*, ECDL '08, pages 371–382, Berlin, Heidelberg, 2008. Springer-Verlag.

[129] Frank McCown and Michael L. Nelson. What happens when facebook is gone? In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 251–254, New York, NY, USA, 2009. ACM.

[130] Martin Klein and Michael L. Nelson. Investigating The Change Of Web Pages' Titles Over Time. Technical Report arXiv:0907.3445, 2009.

[131] Martin Klein, Jeffery L. Shipman, and Michael L. Nelson. Is This A Good Title? In *Proceedings Of The 21st ACM Conference On Hypertext and Hypermedia*, Hypertext '10, pages 3–12, 2010.

[132] Martin Klein and Michael L. Nelson. Inter-search Engine Lexical Signature Performance. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, pages 413–414, New York, NY, USA, 2009. ACM.

[133] Martin Klein and Michael L. Nelson. Evaluating Methods To Rediscover Missing Web Pages From The Web Infrastructure. In *Proceedings Of The 10th Annual Joint Conference On Digital Libraries*, JCDL '10, pages 59–68, New York, NY, USA, 2010. ACM.

[134] Terry L. Harrison and Michael L. Nelson. Just-In-Time Recovery Of Missing Web Pages. In *Proceedings Of The 17th ACM Conference On Hypertext and Hypermedia*, Hypertext '06, pages 145–156, 2006.

[135] Terry L. Harrison. Opal: In Vivo Based Preservation Framework For Locating Lost Web Pages. Master's thesis, Old Dominion University, 2005.

[136] Frank McCown, Joan A. Smith, Michael L. Nelson, and Johan Bollen. Lazy Preservation: Reconstructing Websites By Crawling The Crawlers. In *WIDM '06: Proceedings Of The 8th Annual ACM International Workshop On Web Information and Data Management*, pages 67–74, 2006.

[137] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing User Behavior In Online Social Networks. In *Proceedings Of The*

9th ACM SIGCOMM Conference On Internet Measurement Conference, IMC '09, pages 49–62, New York, NY, USA, 2009. ACM.

[138] Georges Dupret and Mounia Lalmas. Absence Time and User Engagement: Evaluating Ranking Functions. In *Proceedings Of The Sixth ACM International Conference On Web Search and Data Mining*, WSDM '13, pages 173–182, New York, NY, USA, 2013. ACM.

[139] Elad Yom-Tov, Mounia Lalmas, Ricardo A. Baeza-Yates, Georges Dupret, Janette Lehmann, and Pinar Donmez. Measuring Inter-Site Engagement. In *Proceedings Of The 2013 IEEE International Conference On Big Data*, pages 228–236, 2013, 6-9 October 2013, Santa Clara, CA, USA.

[140] Ricardo A. Baeza-Yates and Mounia Lalmas. User Engagement: The Network Effect Matters! In *21st ACM International Conference On Information and Knowledge Management*, pages 1–2, 2012, CIKM'12, Maui, HI, USA, October 29 - November 02, 201.

[141] Elad Yom-Tov, Mounia Lalmas, Georges Dupret, Ricardo A. Baeza-Yates, Pinar Donmez, and Janette Lehmann. The Effect Of Links On Networked User Engagement. In *Proceedings Of The 21st World Wide Web Conference*, pages 641–642, 2012, WWW 2012, Lyon, France, April 16-20, 2012 (Companion Volume).

[142] Janette Lehmann, Mounia Lalmas, Elad Yom-Tov, and Georges Dupret. Models Of User Engagement. In *User Modeling, Adaptation, and Personalization - 20th International Conference, UMAP 2012*, pages 164–175, Montreal, Canada, July 16-20, 2012.

[143] Gilad Mishne and Natalie Glance. Predicting Movie Sales From Blogger Sentiment. In *AAAI Symposium On Computational Approaches To Analysing Weblogs (AAAI-CAAW)*, pages 155–158, 2006.

[144] Kathleen T. Durant and Michael D. Smith. Predicting The Political Sentiment Of Web Log Posts Using Supervised Machine Learning Techniques Coupled With Feature Selection. In *Proceedings Of The 8th Knowledge Discovery On The Web International Conference On Advances In Web Mining and Web Usage Analysis*, WebKDD'06, pages 187–206, Berlin, Heidelberg, 2007. Springer-Verlag.

[145] Kathleen T. Durant and Michael D. Smith. Mining Sentiment Classification From Political Web Logs. In *In Proceedings Of Workshop On Web Mining and Web Usage Analysis Of The 12th ACM SIGKDD International Conference On Knowledge Discovery and Data Mining, WebKDD'06*, 2006.

[146] Onur Kucuktunc, B. Barla Cambazoglu, Ingmar Weber, and Hakan Ferhatosmanoglu. A Large-Scale Sentiment Analysis For Yahoo! Answers. In *Proceedings Of The Fifth ACM International Conference On Web Search and Data Mining - WSDM '12*, page 633, New York, New York, USA, February 2012. ACM Press.

[147] Peter Dodds and Christopher Danforth. Measuring The Happiness Of Large-Scale Written Expression: Songs, Blogs, and Presidents. *Journal Of Happiness Studies*, 11(4):441–456, August 2010.

[148] Eric Gilbert and Karrie Karahalios. Widespread Worry and The Stock Market, ICWSM '10. In *Proceedings Of 4th International AAAI Conference on Weblogs and Social Media*, pages 59–65.

[149] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment. In *Proceedings Of The Fourth International AAAI Conference On Weblogs and Social Media*, pages 178–185, 2010.

[150] D. Gayo-Avello, P. T. Metaxas, and E. Mustafaraj. Limits Of Electoral Predictions Using Twitter. In *5th International AAAI Conference On Weblogs and Social Media (AAAI-ICWSM)*, volume 21, 2011.

[151] F. Abel, Q. Gao, G. J. Houben, and K. Tao. Analyzing User Modeling On Twitter For Personalized News Recommendations. *User Modeling, Adaption and Personalization*, pages 1–12, Springer, 2011.

[152] Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. Arsa: A Sentiment-Aware Model For Predicting Sales Performance Using Blogs. In *Proceedings Of The 30th Annual International ACM SIGIR Conference On Research and Development In Information Retrieval*, SIGIR '07, pages 607–614, New York, NY, USA, 2007. ACM.

[153] Alexander Pak and Patrick Paroubek. Twitter As A Corpus For Sentiment Analysis and Opinion Mining. In *Proceedings Of The Seventh International Conference On Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[154] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment In Twitter Events. *Journal Of The Association For Information Science and Technology (JAIST)*, 62(2):406–418, February 2011.

[155] Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-Supervised Recursive Autoencoders For Predicting Sentiment Distributions. In *Proceedings Of The Conference On Empirical Methods In Natural Language Processing*, EMNLP '11, pages 151–161, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[156] Aditya Mogadala and Vasudeva Varma. Twitter User Behavior Understanding With Mood Transition Prediction. In *Proceedings Of The 2012 Workshop On Data-Driven User Behavioral Modelling and Mining From Social Media*, DUBMMSM '12, pages 31–34, New York, NY, USA, 2012. ACM.

[157] A. Bermingham and A. F. Smeaton. On Using Twitter To Monitor Political Sentiment and Predict Election Results. In *SAAIP: Sentiment Analysis Where AI Meets Psychology, IJCNLP 2011 Workshop*, Chiang-Mai, Thailand, November 2011.

[158] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter Mood Predicts The Stock Market. *Journal Of Computational Science*, 2(1):1–8, March 2011.

[159] G. Miller. Social Scientists Wade Into The Tweet Stream. *Science*, 333(6051):1814–1815, 2011.

[160] J. Bollen, A. Pepe, and H. Mao. Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. Technical Report arXiv:0911.1583, 2009.

[161] Jin Ha Lee and Xiao Hu. Generating Ground Truth For Music Mood Classification Using Mechanical Turk. In *Proceedings Of The 12th ACM/IEEE-CS Joint Conference On Digital Libraries*, JCDL '12, pages 129–138, New York, NY, USA, 2012. ACM.

[162] Xiao Hu and J. Stephen Downie. Improving Mood Classification In Music Digital Libraries By Combining Lyrics and Audio. In *Proceedings Of The 10th Annual Joint Conference On Digital Libraries*, JCDL '10, pages 159–168, New York, NY, USA, 2010. ACM.

[163] G. A. Mishne and M. De Rijke. Capturing Global Mood Levels Using Blog Posts. In *AAAI 2006 Spring Symposium On Computational Approaches To Analysing Weblogs*. Publications of the Universiteit van Amsterdam (Netherlands), AAAI Press, 2006.

[164] Azin Ashkan, Charles L. Clarke, Eugene Agichtein, and Qi Guo. Classifying and Characterizing Query Intent. In *Proceedings Of The 31th European Conference On IR Research On Advances In Information Retrieval*, ECIR '09, pages 578–586, Berlin, Heidelberg, 2009. Springer-Verlag.

[165] Chiung-Hon Leon Lee and Alan Liu. Modeling The Query Intention With Goals. In *Proceedings Of The 19th International Conference On Advanced Information Networking and Applications - Volume 2*, AINA '05, pages 535–540, Washington, DC, USA, 2005. IEEE Computer Society.

[166] Alexander Löser, Wojciech M. Barczynski, and Falk Brauer. What's The Intention Behind Your Query? A Few Observations From A Large Developer Community. In *1st International Workshop on Identity and Reference on the Semantic Web, CEUR-IRSW '08*, volume 422, 2008.

[167] Leif Azzopardi and Maarten de Rijke. Query Intention Acquisition: A Case Study On Automatically Inferring Structured Queries. *Proceedings Of The 6th Dutch-Belgian Information Retrieval Workshop*, pages 3–10, 2006.

[168] Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Cristina González-Caro. The Intention Behind Web Queries. In *Proceedings Of The 13th International Conference On String Processing and Information Retrieval*, SPIRE'06, pages 98–109, Berlin, Heidelberg, 2006. Springer-Verlag.

[169] Eldar Sadikov, Jayant Madhavan, Lu Wang, and Alon Halevy. Clustering Query Refinements By User Intent. In *Proceedings Of The 19th International Conference On World Wide Web*, WWW '10, pages 841–850, New York, NY, USA, 2010. ACM.

[170] Bernard J. Jansen, Danielle L. Booth, and Amanda Spink. Determining The User Intent Of Web Search Engine Queries. In *Proceedings Of The 16th International Conference On World Wide Web*, WWW '07, pages 1149–1150, New York, NY, USA, 2007. ACM.

[171] Xiao Li, Ye-Yi Wang, and Alex Acero. Learning Query Intent From Regularized Click Graphs. In *Proceedings Of The 31st Annual International ACM SIGIR Conference On Research and Development In Information Retrieval - SIGIR '08*, page 339, New York, New York, USA, July 2008. ACM Press.

[172] Botao Hu, Yuchen Zhang, Weizhu Chen, Gang Wang, and Qiang Yang. Characterizing Search Intent Diversity Into Click Models. In *Proceedings Of The 20th International Conference On World Wide Web - WWW '11*, page 17, New York, New York, USA, March 2011. ACM Press.

[173] Xiao Li, Ye-Yi Wang, Dou Shen, and Alex Acero. Learning With Click Graph For Query Intent Classification. *ACM Transactions On Information Systems*, 28(3):12:1–12:20, July 2010.

[174] Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. Intent-Aware Search Result Diversification. In *Proceedings Of The 34th International ACM SIGIR Conference On Research and Development In Information Retrieval*, SIGIR '11, pages 595–604, New York, NY, USA, 2011. ACM.

[175] Na Dai, Xiaoguang Qi, and Brian D. Davison. Enhancing Web Search With Entity Intent. In *Proceedings Of The 20th International Conference Companion On World Wide Web*, WWW '11, pages 29–30, New York, NY, USA, 2011. ACM.

[176] Na Dai, Xiaoguang Qi, and Brian D. Davison. Bridging Link and Query Intent To Enhance Web Search. In *Proceedings Of The 22nd ACM Conference On Hypertext and Hypermedia*, HT '11, pages 17–26, New York, NY, USA, 2011. ACM.

[177] Qi Guo and Eugene Agichtein. Ready To Buy Or Just Browsing?: Detecting Web Searcher Goals From Interaction Data. In *Proceedings Of The 33rd International ACM SIGIR Conference On Research and Development In Information Retrieval*, SIGIR '10, pages 130–137, New York, NY, USA, 2010. ACM.

[178] András Benczúr, István Bíró, Károly Csalogány, and Tamás Sarlós. Web Spam Detection via Commercial Intent Analysis. In *Proceedings of the 3rd International*

Workshop on Adversarial Information Retrieval on the Web, AIRWeb '07, pages 89–92, New York, NY, USA, 2007. ACM.

[179] Min Wu, Robert C. Miller, and Greg Little. Web Wallet: Preventing Phishing Attacks By Revealing User Intentions. In *Proceedings Of The Second Symposium On Usable Privacy and Security*, SOUPS '06, pages 102–113, New York, NY, USA, 2006. ACM.

[180] Ke Zhou, Stewart Whiting, Joemon M. Jose, and Mounia Lalmas. The Impact Of Temporal Intent Variability On Diversity Evaluation. In *Proceedings Of The 35th European Conference On IR Research On Advances In Information Retrieval*, ECIR '13, pages 820–823, 2013.

[181] S. Papadimitriou C. Zhu, H. Kitagawa and C. Faloutsos. A Method Of Detecting Outliers Matching User's Intentions. In *Proceedings of the 15th IEICE Data Engineering Workshop.*, pages 93–96, 2004.

[182] Vinay Jethava, Liliana Calderón-Benavides, Ricardo Baeza-Yates, Chiranjib Bhattacharyya, and Devdatt Dubhashi. Scalable Multi-Dimensional User Intent Identification Using Tree Structured Distributions. In *Proceedings Of The 34th International ACM SIGIR Conference On Research and Development In Information Retrieval*, SIGIR '11, pages 395–404, New York, NY, USA, 2011. ACM.

[183] Yelong Shen, Jun Yan, Shuicheng Yan, Lei Ji, Ning Liu, and Zheng Chen. Sparse Hidden-Dynamics Conditional Random Fields For User Intent Understanding. In *Proceedings Of The 20th International Conference On World Wide Web*, WWW '11, pages 7–16, New York, NY, USA, 2011. ACM.

[184] A. Kathuria, B. J. Jansen, C. Hafernik, and A. Spink. Classifying The User Intent Of Web Queries Using K-Means Clustering. *Internet Research*, 20(5):563–581, 2010.

[185] Eurico Doirado and Carlos Martinho. I Mean It!: Detecting User Intentions To Create Believable Behaviour For Virtual Agents In Games. In *Proceedings Of The 9th International Conference On Autonomous Agents and Multiagent Systems*, AAMAS '10, pages 83–90, volume 1, Richland, SC, 2010. International Foundation for Autonomous Agents and Multiagent Systems.

[186] Angela B. Dalton and Carla S. Ellis. Sensing User Intention and Context For Energy Management. In *Proceedings Of The 9th Conference On Hot Topics In Operating Systems - Volume 9*, HOTOS'03, pages 151–156, Berkeley, CA, USA, 2003. USENIX Association.

[187] Zheng Chen, Fan Lin, Huan Liu, Yin Liu, Wei-Ying Ma, and Liu Wenyin. User Intention Modeling In Web Applications Using Data Mining. *World Wide Web*, 5(3):181–191, November 2002.

[188] Martin Fishbein and Icek Ajzen. *Belief, Attitude, Intention, and Behavior: An Introduction To Theory and Research.* Addison-Wesley, Reading, MA, 1975.

[189] Judith A. Ouellette and Wendy Wood. Habit and Intention In Everyday Life: The Multiple Processes By Which Past Behavior Predicts Future Behavior. *Psychological Bulletin*, 124(1):54–74, 1998.

[190] Kevyn Collins-Thompson, Paul N. Bennett, Fernando Diaz, Craig Macdonald, and Ellen Voorhees. TREC 2014 Web Track. `http://www-personal.umich.edu/~kevynct/trec-web-2014/`, 2014.

[191] Ian Soboroff. Do TREC Web Collections Look Like The Web? *Newsletter, ACM SIGIR Forum*, 36(2):23–31, September 2002.

[192] Wei-Tsen Milly Chiang, Markus Hagenbuchner, and Ah Chung Tsoi. The WT10G Dataset and The Evolution Of The Web. In *Special Interest Tracks and Posters Of The 14th International Conference On World Wide Web*, WWW '05, pages 938–939, New York, NY, USA, 2005. ACM.

[193] Jure Leskovec and Andrej Krevl. Snap Datasets: Stanford Large Network Dataset Collection. `http://snap.stanford.edu/data`, June 2014.

[194] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing User Studies With Mechanical Turk. In *Proceeding Of The Twenty-Sixth Annual CHI Conference On Human Factors In Computing Systems - CHI '08*, page 453, New York, New York, USA, April 2008. ACM Press.

[195] Yuandong Tian and Jun Zhu. Learning From Crowds In The Presence Of Schools Of Thought. In *Proceedings Of The 18th ACM SIGKDD International Conference On Knowledge Discovery and Data Mining*, KDD '12, pages 226–234, New York, NY, USA, 2012. ACM.

[196] Jonathan L. Elsas and Susan T. Dumais. Leveraging Temporal Dynamics Of Document Content In Relevance Ranking. In *Proceedings Of The Third ACM International Conference On Web Search and Data Mining*, WSDM '10, pages 1–10, New York, NY, USA, 2010. ACM.

[197] Robert Kosara and Caroline Ziemkiewicz. Do Mechanical Turks Dream Of Square Pie Charts? In *Proceedings Of The 3rd Beliv'10 Workshop: Beyond Time and Errors: Novel Evaluation Methods For Information Visualization*, BELIV '10, pages 63–70, New York, NY, USA, 2010. ACM.

[198] Jeffrey Heer and Michael Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk To Assess Visualization Design. In *Proceedings Of The Sigchi Conference On Human Factors In Computing Systems*, CHI '10, pages 203–212, New York, NY, USA, 2010. ACM.

[199] Ke Zhou, Ronan Cummins, Mounia Lalmas, and Joemon M. Jose. Which Vertical Search Engines Are Relevant? In *Proceedings Of The 22nd International Conference On World Wide Web*, WWW '13, pages 1557–1568, Rio de Janeiro, Brazil, 2013.

[200] Paul Heymann and Hector Garcia-Molina. Turkalytics. In *Proceedings Of The 20th International Conference On World Wide Web - WWW '11*, page 477, New York, New York, USA, March 2011. ACM Press.

[201] Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. Perspectives On Crowdsourcing Annotations For Natural Language Processing. *Language Resources and Evaluation*, 47(1):9–31, March 2012.

[202] Hany M. SalahEldeen. Losing My Revolution: A Year After The Egyptian Revolution, 10% Of The Social Media Documentation Is Gone. `http://ws-dl.blogspot.com/2012/02/2012-02-11-losing-my-revolution-year.html`, 2012.

[203] Hany M. SalahEldeen and Michael L. Nelson. Resurrecting My Revolution: Using Social Link Neighborhood In Bringing Context To The Disappearing Web. In *Research and Advanced Technology For Digital Libraries - International Conference On Theory and Practice Of Digital Libraries*, TPDL'13, pages 333–345. Springer-Verlag, 2013.

[204] Hany M. SalahEldeen and Michael L. Nelson. Losing My Revolution: How Many Resources Shared On Social Media Have Been Lost? In *Proceedings Of The Second International Conference On Theory and Practice Of Digital Libraries*, TPDL'12, pages 125–137, Berlin, Heidelberg, 2012. Springer-Verlag.

[205] Jaewon Yang and Jure Leskovec. Patterns Of Temporal Variation In Online Media. In *Proceedings Of The Fourth ACM International Conference On Web Search and Data Mining - WSDM '11*, pages 177–186, New York, New York, USA, February 2011. ACM Press.

[206] Alex Nunns and Nadia Idle. Tweets From Tahrir: Egypt's Revolution As It Unfolded, In The Words Of The People Who Made It., 2011, ISBN-10: 1935928457.

[207] Justin F. Brunelle and Michael L. Nelson. An Evaluation Of Caching Policies For Memento Timemaps. In *Proceedings Of The 13th ACM/IEEE-CS Joint Conference On Digital Libraries*, JCDL '13, pages 267–276, New York, NY, USA, 2013. ACM.

[208] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. Boilerplate Detection Using Shallow Text Features. In *Proceedings Of The Third ACM International Conference On Web Search and Data Mining*, WSDM '10, pages 441–450, New York, NY, USA, 2010. ACM.

[209] Scott G. Ainsworth, Ahmed AlSum, Hany M. SalahEldeen, Michele C. Weigle, and Michael L. Nelson. How Much Of The Web Is Archived? In *Proceedings Of The 11th Annual International ACM/IEEE Joint Conference On Digital Libraries*, JCDL '11, pages 133–136, New York, NY, USA, 2011. ACM.

[210] Ziv Bar-Yossef and Maxim Gurevich. Random Sampling From A Search Engine's Index. *Journal Of The ACM*, 55(5):1–74, October 2008.

[211] Ahmed AlSum. *Web Archive Services Framework For Tighter Integration Between The Past and Present Web.* PhD thesis, Old Dominion University Department of Computer Science, 2014.

[212] Yasmin AlNoamany, Michele C. Weigle, and Michael L. Nelson. Characteristics Of Social Media Stories. In *Submitted For Publication*, 2015.

[213] Justin F. Brunelle, Mat Kelly, Hany M. SalahEldeen, Michele C. Weigle, and Michael L. Nelson. Not All Mementos Are Created Equal: Measuring The Impact Of Missing Resources. In *Proceedings Of The 14th Annual International ACM/IEEE Joint Conference On Digital Libraries, Dl 14*, pages 321–330, Sept 2014.

[214] Hany M. SalahEldeen and Michael L. Nelson. Carbon Dating The Web: Estimating The Age Of Web Resources. In *Proceedings Of The 22nd International Conference On World Wide Web Companion*, TempWeb '03, WWW '13 Companion, pages 1075–1082, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[215] Alexander Nwala. Carbon Dating The Web, Version 2.0. `http://ws-dl.blogspot.com/2014/11/2014-11-14-carbon-dating-web-version-20.html`, 2014.

[216] Michael L. Nelson. Memento-Datetime Is Not Last-Modified. `http://ws-dl.blogspot.com/2010/11/2010-11-05-memento-datetime-is-not-last.html`, 2011.

[217] Alexis Rossi. Fixing Broken Links on the Internet. `https://blog.archive.org/2013/10/25/fixing-broken-links/`, 2013.

[218] Frank McCown and Michael L. Nelson. Search Engines and Their Public Interfaces: Which APIs Are The Most Synchronized? In *Proceedings Of The 16th International Conference On World Wide Web, WWW '07*, pages 1197–1198, 2007.

[219] Twitter. Using The Twitter Search API. `https://dev.twitter.com/docs/using-search`, 2013.

[220] Onn Brandman, Junghoo Cho, Hector Garcia-Molina, and Narayanan Shivakumar. Crawler-Friendly Web Servers. *ACM SIGMETRICS Performance Evaluation Review*, 28(2):9–14, September 2000.

[221] Adam Jatowt, Yukiko Kawai, and Katsumi Tanaka. Detecting Age Of Page Content. In *WIDM '07: Proceedings Of The 9th Annual ACM International Workshop On Web Information and Data Management*, pages 137–144, New York, NY, USA, 2007. ACM.

[222] Hany M. SalahEldeen and Michael L. Nelson. Reading The Correct History?: Modeling Temporal Intention In Resource Sharing. In *Proceedings Of The 13th*

ACM/IEEE-CS Joint Conference On Digital Libraries, JCDL '13, pages 257–266, New York, NY, USA, 2013. ACM.

[223] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. In *Proceedings Of The ACL-02 Workshop On Effective Tools and Methodologies For Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[224] G. Holmes, A. Donkin, and I. H. Witten. Weka: A Machine Learning Workbench. In *Proceedings Second Australia and New Zealand Conference On Intelligent Information Systems*, Brisbane, Australia, 1994.

[225] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named Entity Recognition In Tweets: An Experimental Study. In *Proceedings Of The Conference On Empirical Methods In Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[226] Aobo Wang, Tao Chen, and Min-Yen Kan. Re-Tweeting From A Linguistic Perspective. In *Proceedings Of The Second Workshop On Language In Social Media*, LSM '12, pages 46–55, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[227] S. C. Deerwester, S. T. Dumais, G. W. Furnas, R. A. Harshman, T. K. Landauer, K. E. Lochbaum, and L. A. Streeter. Computer Information Retrieval Using Latent Semantic Structure, June 13 1989. US Patent 4,839,853.

[228] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal Of Machine Learning Research*, 3:993–1022, March 2003.

[229] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving Lda Topic Models For Microblogs Via Tweet Pooling and Automatic Labeling. In *Proceedings Of The 36th International ACM SIGIR Conference On Research and Development In Information Retrieval*, SIGIR '13, pages 889–892, New York, NY, USA, 2013. ACM.

[230] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing Twitter and Traditional Media Using Topic

Models. In *Proceedings Of The 33rd European Conference On Advances In Information Retrieval*, ECIR'11, pages 338–349, Berlin, Heidelberg, 2011. Springer-Verlag.

[231] Matthew D. Hoffman, David M. Blei, and Francis R. Bach. Online Learning For Latent Dirichlet Allocation. In *NIPS*, pages 856–864, 2010.

[232] Radim Řehůřek and Petr Sojka. Software Framework For Topic Modelling With Large Corpora. In *Proceedings Of The LREC 2010 Workshop On New Challenges For NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA.

[233] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-Sampling Technique. *Journal Of Artificial Intelligence Research*, 16(1):321–357, June 2002.

[234] Twitter Now Lets You Search For Any Tweet Ever Sent. `http://www.wired.com/2014/11/twitter-now-lets-search-tweet-ever-sent/`, 2014.

[235] Ahmed AlSum and Michael L. Nelson. Thumbnail Summarization Techniques For Web Archives. In *Proceedings Of The 36th European Conference On IR Research On Advances In Information Retrieval*, ECIR '14, pages 299–310, Amsterdam, The Netherlands, 2014. Springer-Verlag.

[236] A. Sarcevic, L. Palen, J. White, K. Starbird, M. Bagdouri, and K. anderson. Beacons Of Hope In Decentralized Coordination: Learning From On-The-Ground Medical Twitterers During The 2010 Haiti Earthquake. In *Proceedings Of The ACM 2012 Conference On Computer Supported Cooperative Work*, pages 47–56, 2012.

[237] Alexander Hanna. Computer-Aided Content Analysis Of Digitally Enabled Movements. *Mobilization: An International Quarterly*, 18(4):367–388, 2013.

VITA

Hany M. SalahEldeen

Department of Computer Science

Old Dominion University

Norfolk, VA 23529

EDUCATION

PhD. in Computer Science, Old Dominion University, USA, 2015

M.S. in Computer Science, Universitat Autònoma de Barcelona, Spain, 2009

B.S. in Computer Systems Engineering, Alexandria University, Egypt, 2008

PROFESSIONAL EXPERIENCE

2012 - 2015 Instructor of Computer Science, Old Dominion University

2014 - 2014 Guest Researcher, School of Computing, National University of Singapore, Singapore

2010 - 2011 Teaching Assistant, Old Dominion University

2010 - 2015 Graduate Research Assistant, Old Dominion University

2009 - 2009 Research Intern, Microsoft Research Advanced Technology Lab, Cairo, Egypt

2011 - 2011 Engineering Intern, Microsoft Inc., Mountain View, CA

2010 - 2010 Software Development Intern, Google GmBH, Zürich, Switzerland

2008 - 2009 Teaching Assistant, Universitat Autònoma de Barcelona, Spain

2006 - 2008 Software Development Engineer and Trainer, eSpace Technologies, Alexandria, Egypt

PUBLICATIONS AND PRESENTATIONS

A complete list is available at `http://www.cs.odu.edu/~hany/HanySalahEldeen.pdf`

PROFESSIONAL SOCIETIES

Association for Computing Machinery (ACM)

Typeset using LATEX.