Detecting, Modeling, and Predicting User Temporal Intention in Social Media

Hany M. SalahEldeen
Old Dominion University
Department of Computer Science
Norfolk, Virginia, USA 23529
hany@cs.odu.edu

ABSTRACT

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. 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. Furthermore, the proposed model will result in two main benefits. First, social media navigation will more closely match the implicit temporal intent of the users. Second, we will leverage the many existing public web archives and the Memento project to integrate the past and current web.

Categories and Subject Descriptors

H.3.5 [**Online Information Services:**]: Data Sharing

General Terms

Design, Experimentation, Human Factors

Keywords

User Intention, Modeling, Social Media, Web Archiving, Digital Preservation

1. INTRODUCTION

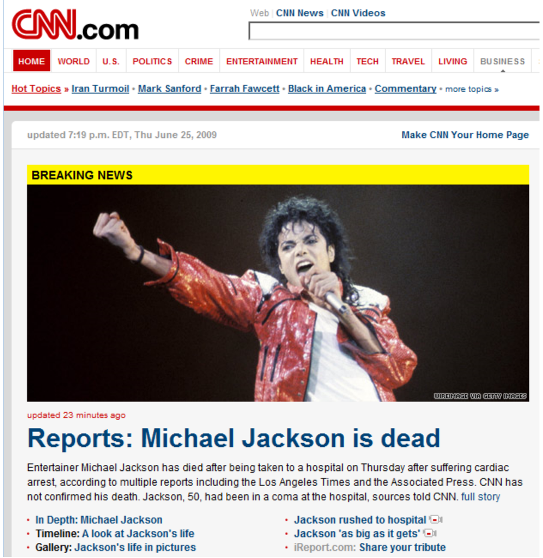
For years content has been posted and shared by users across social networks. In most cases, this content is being shared by reference to the hosting website. For example, someone sees a video on Youtube and shares it on their Facebook account or Twitter by passing a link or URI. The resources pointed to by the link, are prone to change or become lost, making the social post or share completely meaningless. With more than 845 million Facebook users at the end of 2011 [1] and over 140 million tweets sent daily in 2011 [4] users can post photos, videos, personal opinions and report incidents as they happen. Many of the posts and tweets are about quotidian events and their preservation is debatable. However, some of the posts and events are about culturally important events whose preservation is less controversial.

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, getting cut off and information exposure. 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 [24]. 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 Bit.ly, 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 Bit.ly in 2009 and then recently replaced it by its own shortening service t.co. 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, ease of spread, posting, and sharing in social networks.

On the one hand, content has been posted and shared between users in social networks for years. This content is disappearing with no possibility of recovery. On the other hand, people share content and shortened URLs for various reasons, some of which could be malicious and harmful. Also upon sharing a resource that changes frequently, the author might intend to share a certain version which could change completely when an author's friend on the social network resolves this resource.

To illustrate this problem, imagine a user named Jeff. Jeff has a Twitter account and he is relatively active and many of his friends follow him. On the 25th of June 2009, he woke up, checked cnn.com and found that Michael Jackson was dead (Figure 1). He was shocked and posted this bad news on his Twitter account (Figure 2).

His friend Jenny was in Hawaii on a vacation and decided to leave her phone and laptop back in her apartment to stay off the grid. When she came back after a month on July


CNN.com
Web | CNN News | CNN Videos
HOME WORLD U.S. POLITICS CRIME ENTERTAINMENT HEALTH TECH TRAVEL LIVING BUSINESS
Hot Topics » Iran Turmoil • Mark Sanford • Farrah Fawcett • Black in America • Commentary • more topics »
Make CNN Your Home Page
WIREIMAGE VIA GETTY IMAGES

updated 7:19 p.m. EDT, Thu June 25, 2009

BREAKING NEWS

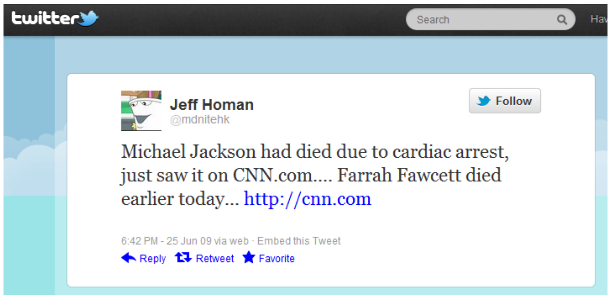
updated 23 minutes ago

Reports: Michael Jackson is dead

Entertainer Michael Jackson has died after being taken to a hospital on Thursday after suffering cardiac arrest, according to multiple reports including the Los Angeles Times and the Associated Press. CNN has not confirmed his death. Jackson, 50, had been in a coma at the hospital, sources told CNN. full story

- In Depth: Michael Jackson
- Timeline: A look at Jackson's life
- Gallery: Jackson's life in pictures
- Jackson rushed to hospital
- Jackson 'as big as it gets'
- iReport.com: Share your tribute

Figure 1: CNN Front page on June 25th 2009.



Jeff Homan
@mdnitehk

Michael Jackson had died due to cardiac arrest, just saw it on CNN.com.... Farrah Fawcett died earlier today... http://cnn.com

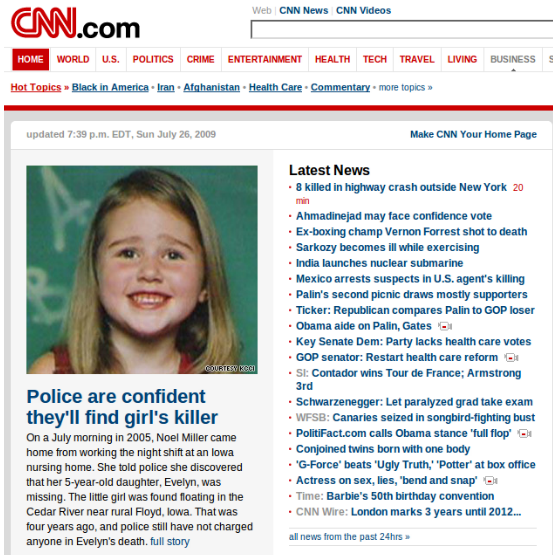
6:42 PM - 25 Jun 09 via web · Embed this Tweet

Figure 2: Jeff's tweet on the 25th of June 2009

the 26th, she started checking her emails and her friends' facebook statuses and tweets. Upon reading Jeff's tweet, she was also shocked and opened the link he embedded in the tweet referencing cnn.com and found it completely different (Figure 3). She thought he was playing a prank on her and she got mad at him.

This is a simple example that illustrates the problem of Jeff's temporal intention. He intended to share CNN.com as it existed on June 25th 2009 at 7pm, regardless of when Jenny clicked on the link. Knowing that most of the content posted or shared on social networks is done by reference to the hosting website, this content is vulnerable to change or loss. This incoherence between the context of the post and the shared resource associated with this post makes it prone to be completely unuseful or even worse, unrecoverable.

To explain what happened in the previous example, Figure 5 shows the incidents related to this event across time. The resource (CNN.com) was updated at time t_0; a snapshot was taken and preserved in the archives at time t_1; and Jeff posted his tweet at time t_2. At time t_3, the content of CNN's front page changed and shortly after that, a snapshot was taken into the archives at time t_4. Following that at times t_5 to t_n the content continued to change and some of these changes were captured in the archives until the evening of July 26th when Jenny checked the link Jeff posted and found completely different content.


CNN.com
Web | CNN News | CNN Videos
HOME WORLD U.S. POLITICS CRIME ENTERTAINMENT HEALTH TECH TRAVEL LIVING BUSINESS
Hot Topics » Black in America • Iran • Afghanistan • Health Care • Commentary • more topics »
Make CNN Your Home Page
COURTESY KCCI

updated 7:39 p.m. EDT, Sun July 26, 2009

Police are confident they'll find girl's killer

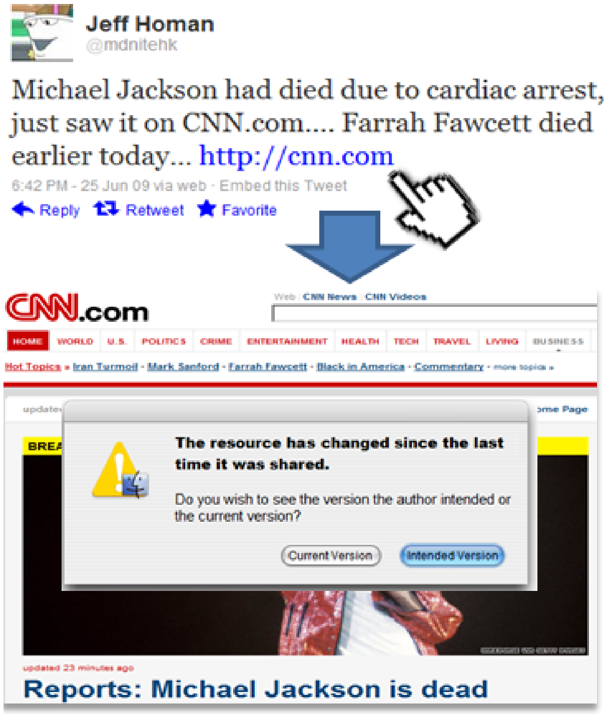
On a July morning in 2005, Noel Miller came home from working the night shift at an Iowa nursing home. She told police she discovered that her 5-year-old daughter, Evelyn, was missing. The little girl was found floating in the Cedar River near rural Floyd, Iowa. That was four years ago, and police still have not charged anyone in Evelyn's death. full story

Latest News

- 8 killed in highway crash outside New York 20 min
- Ahmadinejad may face confidence vote
- Ex-boxing champ Vernon Forrest shot to death
- Sarkozy becomes ill while exercising
- India launches nuclear submarine
- Mexico arrests suspects in U.S. agent's killing
- Palin's second picnic draws mostly supporters
- Ticker: Republican compares Palin to GOP loser
- Obama aide on Palin, Gates
- Key Senate Dem: Party lacks health care votes
- GOP senator: Restart health care reform
- SI: Contador wins Tour de France; Armstrong 3rd
- Schwarzenegger: Let paralyzed grad take exam
- WFSB: Canaries seized in songbird-fighting bust
- PolitiFact.com calls Obama stance 'full flop'
- Conjoined twins born with one body
- 'G-Force' beats 'Ugly Truth,' 'Potter' at box office
- Actress on sex, lies, 'bend and snap'
- Time: Barbie's 50th birthday convention
- CNN Wire: London marks 3 years until 2012...

all news from the past 24hrs »

Figure 3: CNN Front page on July 26th 2009.


Reply
Retweet
Favorite
CNN.com
Web | CNN News | CNN Videos
HOME WORLD U.S. POLITICS CRIME ENTERTAINMENT HEALTH TECH TRAVEL LIVING BUSINESS
Hot Topics » Iran Turmoil • Mark Sanford • Farrah Fawcett • Black in America • Commentary • more topics »
BREA
Current Version
Intended Version

Jeff Homan
@mdnitehk

Michael Jackson had died due to cardiac arrest, just saw it on CNN.com.... Farrah Fawcett died earlier today... http://cnn.com

6:42 PM - 25 Jun 09 via web · Embed this Tweet

The resource has changed since the last time it was shared.

Do you wish to see the version the author intended or the current version?

updated 23 minutes ago

Reports: Michael Jackson is dead

Figure 4: A mockup of the proposed model detecting the change in the resource.

To solve this problem, we introduce the concept of user's temporal intention upon posting or sharing content in online social networks. When users author posts which have embedded or shared resources, we want to detect and estimate which version of each resource they intended to share and which version of the resource they expect readers of their post to view. The state of the resource at the time of sharing or the most recent version that is available of the resource at the time of the reader "clicking". We propose investigating this temporal intention and the factors affecting it. Furthermore, we propose defining and training a parametric model that is able to detect and estimate this temporal intention and can be used in providing a seamless navigational time travel of the posted resources. This temporal prediction would be calculated even before the reader clicks on the URI (or short URI), linking the post and the shared resource. Illustrated along with information about the existence of other copies of the resource, the calculated prediction will give the reader an educated estimation as to which version he/she should read. This pre-clicking information will also act as the first line of defense in case of spam links and page hijacking which could happen as the result of providing the opaque, shortened URI instead of the one posted originally. For example, when Jenny clicks on the link in Jeff's tweet, the model will estimate that this resource is no longer about the context of the tweet and it has been changed or lost since then. It will calculate the most appropriate version and will extract it from the public web archives and notify Jenny with this recommendation (Figure 4).

2. RESEARCH GOALS

Given the problem of inconsistency of the social content published across the times of resolving the associated links (and in some cases, the total loss of this content) we highlight our research question as: Can we estimate the users' intention at the time of posting and reading to predict and maintain temporal consistency?

We can divide the process of sharing resources in the social media into two phases: (1) posting phase, and (2) the reading phase. In each phase, the user has an implicit intention for the version of the shared resource that he/she expects to post or read. Our research aims to understand and analyze this implicit intention and highlight the explicit intention at the time of posting (by asking the author which version does he intend to share) and at the time of reading (by asking the reader which version of the resource does he expect to read). Given that, we set our research goals for this proposal as follows:

- Detect the temporal intention of the author upon sharing a resource in a social network (i.e., did the author intend to share the state of the resource at the moment of sharing or will any version of the resource be sufficient?).
- Model this temporal intention as a function of time, the nature of the resource, its frequency of change, the context of the shared resource, the nature of the social network, and other factors under investigation.
- Implement the model and utilize it in predicting this temporal behavior for resources that have been posted and shared before in order to minimize in consistency and loss of the resources.
- Enhance the implementation, using a training cycle, to be able to automatically preserve vulnerable social content that is prone to change or loss.
- Create a reference implementation of the framework that provides a smooth temporal navigation of the web through current and preserved versions of the resources that are shared on the social network.

3. RELATED WORK

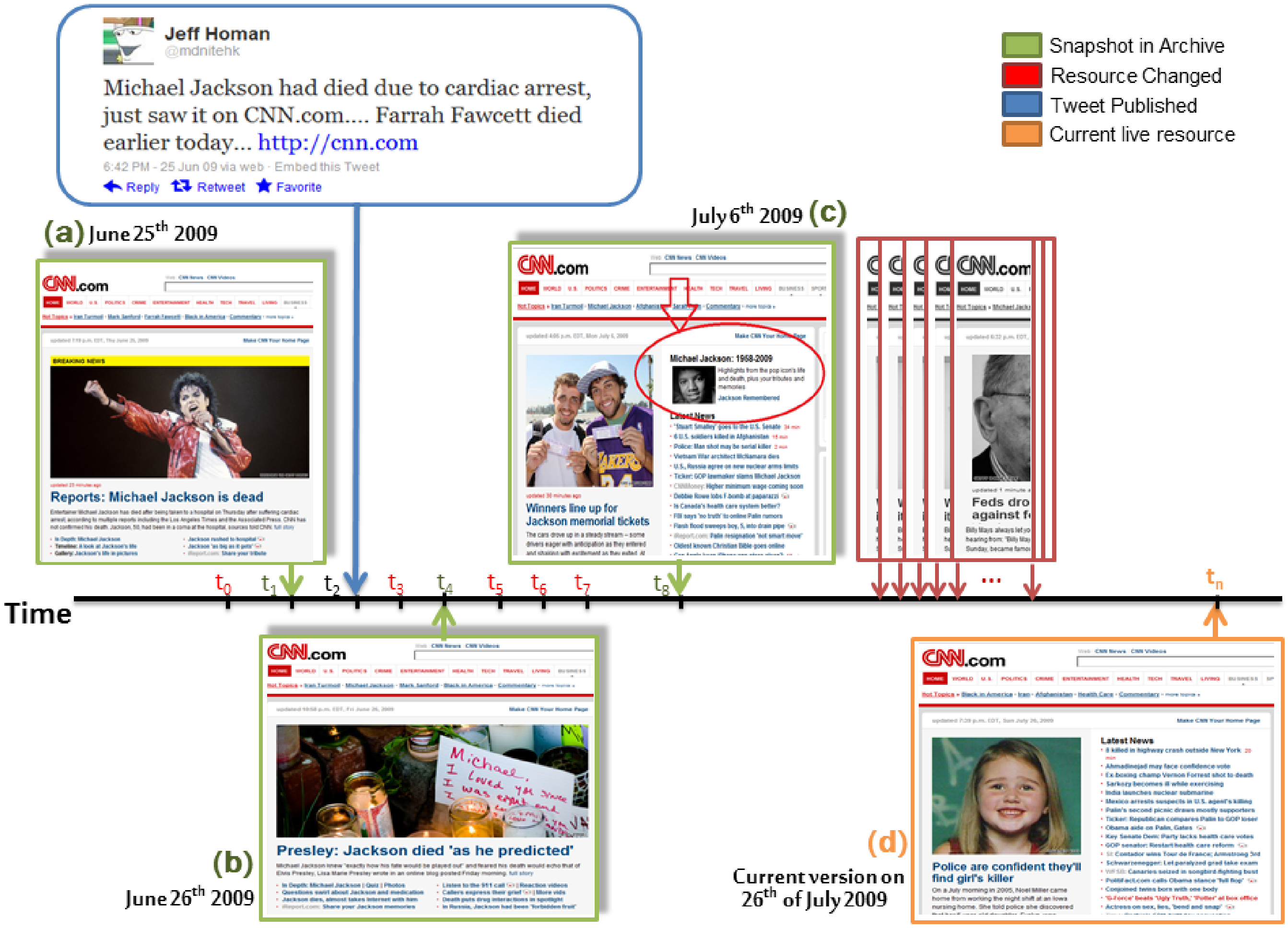
User intent has been studied, analyzed and predicted in several works in the past decade. These works span multiple fields ranging from psychology, sociology, computer engineering, to computer science. Focusing on the latter field, user intention has been tackled from different angles. On the one hand, researchers have studied and analyzed the user intent behind queries in web search [7, 20, 21, 8, 9]. On the other hand, Na Dai et al. proposed classifying the intent expressed by web content creators and classified it as navigational or informational [14]. 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 [13]. User intention has also been studied extensively in the commercial field. Qi Guo et al. 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 [16]. Furthermore, commercial intent analysis was used in web spam detection and resulted in improving the spam classification by 3% [11].

In many cases, sentiment and intent go hand-in-hand in analyzing social networks interactions and posts in the blogosphere. Mishne et al. analyzed the sentiment in weblog posts to predict the movie sales [25]. 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 [15]. Bollen et al. used sentiment analysis in analyzing more than 10 million tweets to predict the Dow Jones Industrial Average (DJIA) with an accuracy of 87.6% in predicting the daily up and down changes in the closing values [12].

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 sentiment analysis. It has not been applied to the temporal intention of users and the bridge between the current and past web.

We analyzed the literature that examined the the persistence of shared resources and web content in general. Nelson and Allen studied the persistence of objects in a digital library and found that within just over a year, 3% of the sample they collected had appeared as no longer be available [26]. Sanderson et al. analyzed the persistence and availability of web resources referenced from papers in scholarly repositories and found that 28% of these resources have been lost [29]. Memento [30] is a collection of HTTP extensions that enables uniform, inter-archive access. McCown et al. examined the factors affecting reconstructing websites (using caches and archives) and found that PageRank, Age, and the number of hops from the top-level of the site were most influential [22].

As for URL shortening, this field is new and to our knowledge no other study, beside the work done by Antoniades et al. [6], have tackled this field. In that study, they argue that



Jeff Homan
@mdnitehk

Michael Jackson had died due to cardiac arrest, just saw it on CNN.com.... Farrah Fawcett died earlier today... http://cnn.com

6:42 PM - 25 Jun 09 via web · Embed this Tweet

Reply
Retweet
Favorite
🟩 Snapshot in Archive
🟥 Resource Changed
🟦 Tweet Published
🟧 Current live resource
(a) June 25th 2009
July 6th 2009 (c)
Reports: Michael Jackson is dead
Michael Jackson: 1968-2009
Winners line up for Jackson memorial tickets
Feds dro against f
Time
t0
t1
t2
t3
t4
t5
t6
t7
t8
...
tn
(b)
June 26th 2009
Presley: Jackson died 'as he predicted'
(d)
Current version on 26th of July 2009
Police are confident they'll find girl's killer
Latest News

Figure 5: Explanation of incidents across time.

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.

Numerous other studies informed our planning due to their relevance to our research. For example, the effect of the author of the content shared and the concept of popularity [31], the general intention of microblogging and tweeting [17, 19], how social networks grow and their evolution in time [23].

4. PRELIMINARY WORK

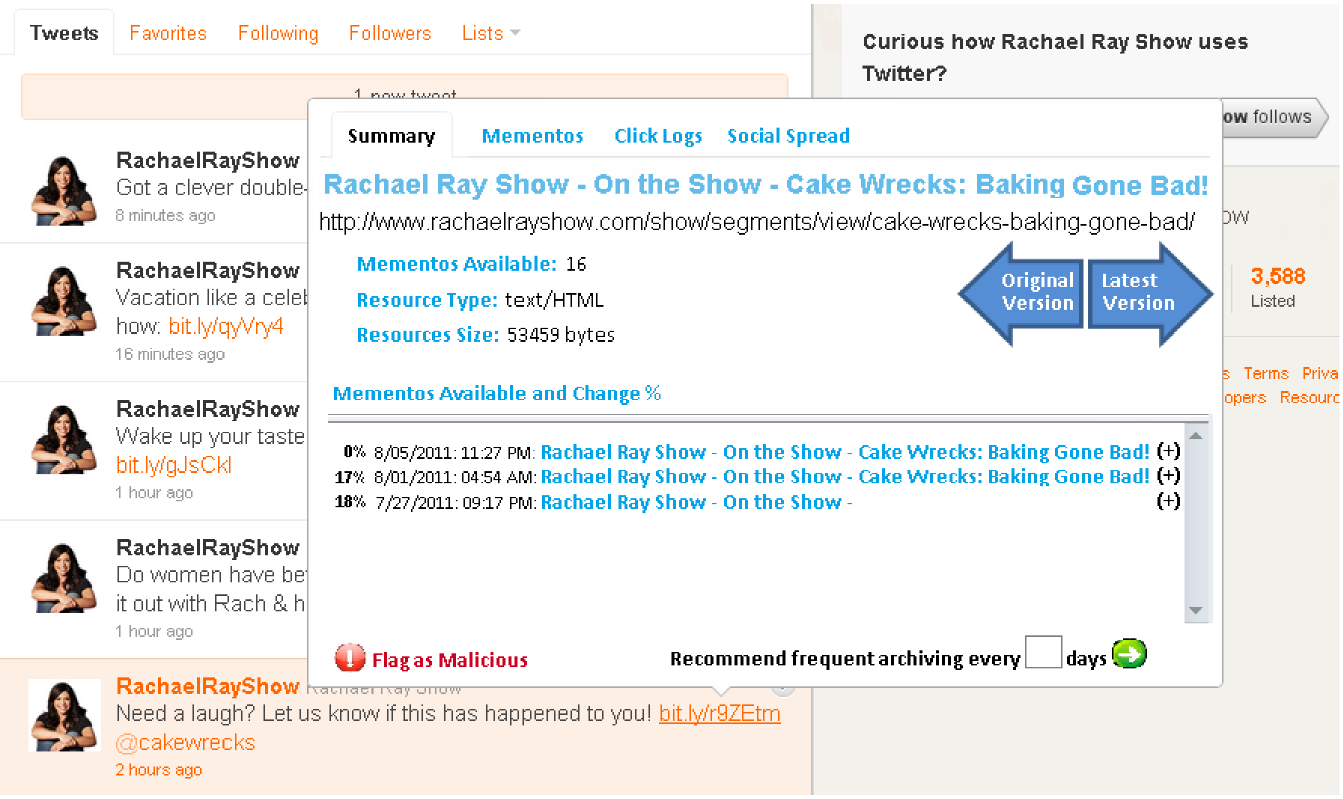
In this section we describe our work to date. We considered several aspects in our analysis, there was a resource nature aspect, a preservation aspect, a social network aspect and finally a human aspect. In the next subsections we examine each of those aspects and explain how they are shaping our understanding of the problem.

4.1 Estimating Web Archiving Coverage

In order to estimate the ability of the web archives to provide versions of the resource posted or shared in social networks we had to estimate the archival coverage. To address this, we sampled 4000 URIs and measured their coverage in the public web archives and the density of this coverage if it exists [5]. We sampled URIs from DMOZ, Delicious, Bitly, and search engine indices and measuring number of archived copies available in various public web archives. The search engine indices were randomly sampled using the technique of Bar-Yossef which attempts to remove the search engine bias towards "popular" resources [10]. The results indicate that the source of the URI plays an important role in how much it is archived. We found that, according to the URI source, the archived percentage varies from 16% to 79%.

4.2 Estimating Social Media Content Loss

After estimating how much of the web is archived we decided to give a closer look to the content shared on social networks. We wanted to answer the question: How much of the social content shared in social networks have been lost and how much can be restored from archives [27]. Also we wanted to examine if there was a relation between the content loss and time. After several experiments, we successfully estimated the content lost of social media and the content archived both as a function of age (time from first share). We extracted social media content that was posted and shared in relation to six public events in the span of three years and found a nearly linear relationship between time of sharing



1 new tweet

RachaelRayShow
Got a clever double-
8 minutes ago

RachaelRayShow
Vacation like a celeb
how: bit.ly/qyVry4
16 minutes ago

RachaelRayShow
Wake up your taste
bit.ly/gJsCkl
1 hour ago

RachaelRayShow
Do women have be
it out with Rach & h
1 hour ago

RachaelRayShow Rachael Ray Show
Need a laugh? Let us know if this has happened to you! bit.ly/r9ZEtm @cakewrecks
2 hours ago

Rachael Ray Show - On the Show - Cake Wrecks: Baking Gone Bad!

http://www.rachaelrayshow.com/show/segments/view/cake-wrecks-baking-gone-bad/

Mementos Available: 16

Resource Type: text/HTML

Resources Size: 53459 bytes

Mementos Available and Change %

```
 0%  8/05/2011: 11:27 PM: Rachael Ray Show - On the Show - Cake Wrecks: Baking Gone Bad! (+)
17%  8/01/2011: 04:54 AM: Rachael Ray Show - On the Show - Cake Wrecks: Baking Gone Bad! (+)
18%  7/27/2011: 09:17 PM: Rachael Ray Show - On the Show -                              (+)
```

Recommend frequent archiving every □ days

Figure 6: Application: Wayback Shortner.

of the resource and the percentage lost. While slightly less linear, we also found a relationship between the archiving coverage of the resource and the time of sharing. From this model we conclude that after the first year of publishing, nearly 11% of this content will be lost and after that we will continue on losing 0.02% per day [28].

4.3 Shortened URIs Analysis

After examining the preservation aspect and a part of the social network aspect we decide to perform experiments that will enable us to have a better understanding of the nature of the resource. We collected a list of shortened URIs that were posted/shared in the social network at one point of time. We started by getting freshly posted/shortened URIs and started a longitudinal study examining each shortened URI collecting click logs, rate of change, rate of spread (re-sharing in the social network) and number of backlinks. Several static properties were collected as well, like the depth of the resource, length of the long original URL, estimated age of the target resource, shortening date, and number of mementos in the archives. Also the persistence of the shortened URI itself will be tested as in some cases both the social post and the shared resource persist but the connecting shortened URI ceases to function as a result of temporary or permanent malfunction of the URL shortening provider for example, the Tr.im URL shortener shut down in 2009. Due to the significance of the problem, the Internet Archive created an independent service called 301Works to provide protection for every day users of short URL services to provide permanence to their mappings[1]. Thus, the shortened URIs were examined to identify their characteristics and how they will provide us with deeper understanding of the shared resources.

4.4 Application: Archive Shortener

We decided to create a simple version of what the application will look like. A plugin was built to integrate with the browser and provide a cross platform, easy to install and light weight environment. It was developed in Javascript and written as a user script which could be installed natively on Google Chrome as it has an embedded support for userscripts, in Firefox (with Grease Monkey [2]), in Opera 8+ as it also has embedded support for userscripts, in Internet Explorer (have IEPro7, and Grease Monkey for IE), and also in Safari 5+ with Grease Kit. This application simply provides an intermediate step between clicking on the resource's shortened URI and opening that resource. By hovering on a shortened URI in a web page, the plugin displays a window describing the title of the resolved form of the resource, a count of how many mementos or snapshots of this resource in the public archives, and the type of the resource. A snapshot of this application is shown in Figure 6.

4.5 User Intention Analysis

Unfortunately, there is no gold standard for determining the temporal intent for social media users. To get ground truth

[1] http://www.301works.org

data for future experiments, we used Amazon's Mechanical Turk[2]. We first tried to estimate the categories that users might divide the resource, original version intention, or latest version intention, or the timeless intention version where there any state the resource is in will be sufficient (e.g., sending your friend a wikipedia page), or finally there is not enough information to decide. Test subjects tend to prefer seeing a snapshot of the latest version side by side to the original version upon deciding which version that best describes a resource in tweet or a post. We are still investigating what other features should be extracted from this experiment.

5. PROPOSED WORK

To fulfill the assigned goals our research will be divided into several phases as explained in the next subsections. Finally, in Table 1 we show the plan towards the completion of my research and thesis. As an added evaluation measure of success in each phase we plan to produce at least one publication by the end of each phase.

5.1 Data Gathering

In this phase we will gather a large dataset of tweets or Facebook posts from different venues and subjects each having an embedded resource (or shortened URI). We have several bigger datasets in our possession to extract this dataset like Stanford's SNAP project tweet dataset [3] or Archeif.org Twitter and Facebook collections. In order to extract the temporal intention behind the publishing each of those social posts we need to present them to a group of human subjects and record their observations. As discussed in section 4.5 above, we will use Amazon's Mechanical Turk in utilizing a large number of reviewers to read the posted tweet in the collection along with its associated resource and estimate the temporal intention that they think the author had at the time of sharing. For example, in the case of Michael Jackson's example (Section 1), Mechanical Turk reviewers will read the tweet, an archived copy of the shared resource, and the current version of the shared resource and can estimate that Jeff (the author) intended to share the snapshot of the CNN front page at June 25th 2009 at 7pm and not the current state of the page. This experiment should be repeated from 5 to 10 times per tweet to test the cross-rater agreement.

5.2 Feature Extraction

In this phase we analyze each resource and collect its clicklog data. Also a group of other features are extracted from the resource like: lexical signatures [18], back links, estimated age, number of times of sharing, depth of the resource, and its frequency of change. Also in this phase we download and analyze all the available versions of the resource in the public archives and estimate the closest version to the creation date of the post.

5.3 Modeling

In this phase we utilize the extracted features to train a classifier and pick the optimal collection of features that provides highest accuracy of classification with the least complexity. We divide our dataset into two subsets, the training set and the testing set. The first part of the modeling phase is to train a simple classifier and use a Naïve Bayes or SVM classifier. After that we plan to extend this model and add more features and remove irrelevant ones. The second part of this modeling phase is to estimate a parametric model from the trained classifier. We will use the ground truth data from Section 4.5 to evaluate our results. Another evaluation criterion will be if we can classify shared resources in real-time.

5.4 Application: Prediction and Preservation

In this phase we implement the model in a web browser plugin. This plugin will analyze shortened URIs and embedded resources within the social media content (e.g., tweets, Facebook posts, blog posts) and perform a dynamic real-time estimation of the suitable version intended for each post. After performing this estimation and upon clicking on the embedded resource in each post, the plugin gives the user an option of opening the current version of the resource or the estimated intended one. It provides a probability estimate for the users and at the same time takes their feedback of their decisions and submit it to the server to be used in the re-learning process of the model. Also the plugin provides the facility of pushing a snapshot of the state of the resource to the public archives upon request of the user. Also upon posting or sharing a resource the plugin estimates the resource's vulnerability of change or loss and provide a recommendation to the author to preserve a current snapshot of the resource and incorporate a shortened URI to the archived snapshot to maintain consistency. This step could be set to run automatically upon each post.

As a final result, this application and the integrated model will provide a time travelling experience, based on user intention, for the user upon reading social media content enabling the reader to seamlessly navigate content through time. Going back to our original Michael Jackson's scenario, when Jenny clicks on the associated link the plugin will provide her with an estimation of change in the resource and its closest archived version to the tweet creation date as we mentioned earlier in Figure 4.

6. CONCLUSIONS

This project addresses the problem of inconsistency between the state of a resource when an author creates a link and the state of the resource when a reader followed the link. This is especially a problem in social media, where sharing resources and providing commentary about the shared resources is a common use case. Users have always had a temporal intent with regard to the resources they share, but until now it has always been implicit. Using web archives, personalized shortened URIs, and analysis of the social media, target resources, and usage patterns, we can make this intent explicit and provide a seamless integration of the current and past web in order to faithfully render the temporal intention of users.

7. ACKNOWLEDGMENTS

I would like to thank my advisor Dr. Michael Nelson for his efforts, guidance, support and understanding. I would like to also extend my thanks and gratitude to the reviewers and readers of this extended abstract, I really appreciate

[2] https://www.mturk.com

Time Frame	Phase Description	Evaluation
Present-8/2012	User Intention Analysis Experiments.	Success measured by having a good large dataset of user intention.
9/2012-12/2012	Candidacy proposal.	Submit my candidacy proposal and defend it.
12/2012-3/2012	Feature Extraction and Training Experiments.	Getting the first phase of the model done and obtain percentage of success upon testing against the dataset.
3/2013-8/2013	Creating the full User Intention Model.	Utilizing the best combination of features to produce the highest accuracy.
8/2013-12/2013	Transforming the classifier Model to a Parametric Model.	Estimate the parametric model that best describes the classifier model trained.
12/2014-2/2014	Creating the first version of the application utilizing the model with self preserving upon shortening.	Publishing the Application and creating a feedback cycle for training the model from the actions of the users.
2/2014-5/2014	Writing the thesis and the Defense.	Finalizing all experiments, arranging publications and finishing the chapters.
6/2014	PhD Defense.	Graduating.

Table 1: The schedule of the research plan.

your feedback and guidance. This work was supported in part by the Library of Congress and NSF IIS-1009392.

8. REFERENCES

[1] Facebook official fact sheet. http://newsroom.fb.com/content/default.aspx?NewsAreaId=22, 2012. [Online; accessed 13-Apr-2012].

[2] Grease Monkey Mozilla Firefox addon. https://addons.mozilla.org/en-US/firefox/addon/greasemonkey/, 2012. [Online; accessed 13-Apr-2012].

[3] Stanford's SNAP Project Twitter dataset. http://snap.stanford.edu/data/twitter7.html, 2012. [Online; accessed 13-Apr-2012].

[4] Twitter numbers. http://blog.Twitter.com/2011/03/numbers.html, 2012. [Online; accessed 13-Apr-2012].

[5] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. 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, 2011.

[6] D. Antoniades, I. Polakis, G. Kontaxis, E. Athanasopoulos, S. Ioannidis, E. P. Markatos, and T. 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.

[7] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. 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.

[8] L. Azzopardi and M. de Rijke. Query intention acquisition: A case study on automatically inferring structured queries. In *Proceedings DIR-2006*, 2006.

[9] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro. The intention behind web queries. In F. Crestani, P. Ferragina, and M. Sanderson, editors, *String Processing and Information Retrieval*, volume 4209 of *Lecture Notes in Computer Science*, pages 98–109. Springer Berlin / Heidelberg, 2006. 10.1007/11880561_9.

[10] Z. Bar-Yossef and M. Gurevich. Random sampling from a search engine's index. In *Proceedings of the 15th international conference on World Wide Web*, WWW '06, pages 367–376, New York, NY, USA, 2006. ACM.

[11] A. Benczúr, I. Bíró, K. Csalogány, and T. 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.

[12] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8, 2011.

[13] N. Dai, X. Qi, and B. 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.

[14] N. Dai, X. Qi, and B. 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.

[15] K. Durant and M. Smith. Predicting the political sentiment of web log posts using supervised machine learning techniques coupled with feature selection. In O. Nasraoui, M. Spiliopoulou, J. Srivastava,

B. Mobasher, and B. Masand, editors, *Advances in Web Mining and Web Usage Analysis*, volume 4811 of *Lecture Notes in Computer Science*, pages 187–206. Springer Berlin / Heidelberg, 2007. 10.1007/978-3-540-77485-3_11.

[16] Q. Guo and E. 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.

[17] A. Java, X. Song, T. Finin, and B. 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.

[18] M. Klein and M. Nelson. Revisiting lexical signatures to (re-)discover web pages. In *Research and Advanced Technology for Digital Libraries*, volume 5173 of *Lecture Notes in Computer Science*, pages 371–382. Springer Berlin / Heidelberg, 2008.

[19] H. Kwak, C. Lee, H. Park, and S. 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.

[20] C.-H. L. Lee and A. 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.

[21] A. Löser, W. M. Barczynski, and F. Brauer. What's the intention behind your query? a few observations from a large developer community. In *Proceedings of the 1st IRSW2008 International Workshop on Identity and Reference on the Semantic Web*, 2008.

[22] F. McCown, N. Diawara, and M. L. Nelson. Factors affecting website reconstruction from the web infrastructure. In *JCDL '07: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 39–48, 2007.

[23] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes. We know who you followed last summer: inferring social link creation times in twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 517–526, New York, NY, USA, 2011. ACM.

[24] N. Megiddo and K. S. McCurley. Efficient retrieval of uniform resource locators, US Patent US 6957224 2005. http://www.google.com/patents/US6957224.

[25] G. Mishne. Predicting movie sales from blogger sentiment. In *In AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, 2006.

[26] M. L. Nelson and B. D. Allen. Object persistence and availability in digital libraries. *D-Lib Magazine*, 8(1), 2002.

[27] H. SalahEldeen and M. L. Nelson. 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.

[28] H. SalahEldeen and M. L. Nelson. Losing my revolution: How resources shared in social media have been lost? In *Accepted for publication at TPDL'12*, 2012.

[29] R. Sanderson, M. Phillips, and H. Van de Sompel. Analyzing the persistence of referenced web resources with Memento. *CoRR*, abs/1105.3459, 2011.

[30] H. Van de Sompel, M. L. Nelson, R. Sanderson, L. Balakireva, S. Ainsworth, and H. Shankar. Memento: Time travel for the web. *CoRR*, abs/0911.1112, 2009.

[31] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 705–714, New York, NY, USA, 2011. ACM.