Building and archiving event web collections: A focused crawler approach

Mohamed M. G. Farag
Dept. of Computer Science, Virginia Tech
Blacksburg, VA 24061 USA
mmagdy@vt.edu

Edward A. Fox
Dept. of Computer Science, Virginia Tech
Blacksburg, VA 24061 USA
fox@vt.edu

ABSTRACT

In this paper, we present a new approach for building and archiving web collections about events. Our approach combines the traditional focused crawling technique with event modeling and representation.

1. INTRODUCTION

In the NSF funded Integrated Digital Event Archiving and Library (IDEAL) project [1] we are building an integrated system to collect, archive, organize, analyze (identify topics, summarize), access (search, browse), and visualize webpages and tweets about real world events (disasters as well as community and government activities). Event archiving is different from Domain/Site-based or Topic-based archiving. The first involves archiving a specific domain/website with all or some of the underlying subdomains/structure. The second covers a given number of webpages related to a user-defined topic.

We have identified and employed three approaches for archiving webpages about events:

1. Manual curation by domain experts, librarians/archivists, and government agencies (High quality – time consuming).
2. Social media-based (crowd sourcing) curation by extracting, retrieving, and archiving URLs from tweet collections about an event (Low quality – time saving). See tweet collections listed on our website [1].
3. Crawling the Web using a focused crawling approach tailored to events (acceptable quality and time).

2. Manual Curation

We have created ~60 collections. These collections are about disaster events--bombings, earthquakes, hurricanes, plane crashes, shootings, floods, fires--and were manually curated and archived using the Archive-it service (https://archive-it.org/organizations/156). Table 1 shows a sample of web collections built using the first approach, manual curation.

3. Social media-based curation

We created more than 600 tweet collections with ~1 billion tweets. From a tweet collection, we extracted the URLs in the tweets, fetched the corresponding webpages, archived just those webpages, and extracted and indexed the text of those webpages.

.

Table 1. Sample of manually curated web collections

Collection Name	No. of Seeds
Alabama University Shooting	116
April 16 Archive	88
Chile Earthquake	19
Nevada air race crash	64
China Floods	60
Encephalitis (India)	59
Hurricane Irene	70

Table 2. Sample of tweet collections

Collection	Keywords/Hashtags	# Tweets
Hurricane Sandy	hurricane sandy	3,219,383
Ebola	#ebola	1,855,680
Ferguson shooting	#Ferguson	1,580,479
Thanksgiving	#Thanksgiving	214,888
AirAsia Plane Crash	#QZ8501	174,353
Charlie Hebdo shooting	#CharlieHebdo	451,009
Iran Talks	#IranTalks	117,966

The tweet/webpage collections are of two types: Disaster events (shootings, earthquakes, plane crashes, hurricanes, bombings, terrorism, floods, and fire) and Community and political events. Table 2 shows a sample of the tweet collections. For a full list please check: http://hadoop.dlib.vt.edu:81/twitter/. Figure 1 shows our process of creating web collection from a tweet collection.

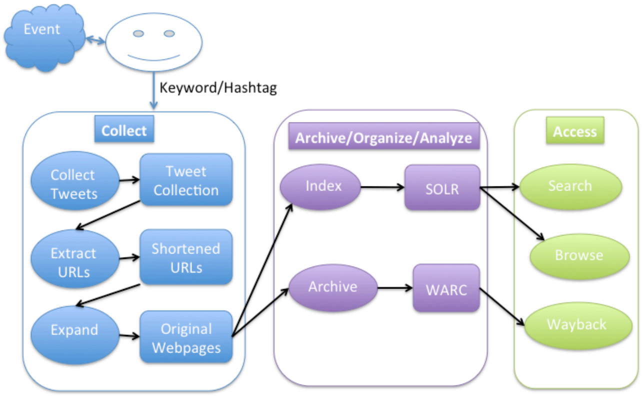

Figure 1. Steps for creating web collections from tweet collections

4. Focused crawler-based curation

This approach [2] aims to maintain a balance between producing high quality event collections and reducing the time/resources needed for collection building. A Curator selects high quality seed URLs and uses the event focused crawler (EFC) to retrieve webpages that are highly similar to those with the seed URLs. The curator can configure EFC to adjust the number of webpages retrieved and the quality of retrieved webpages (similarity threshold). Figure 2 shows the architecture of an event focused crawler. The dotted part highlights the event modeling step and its role in the focused crawling approach.

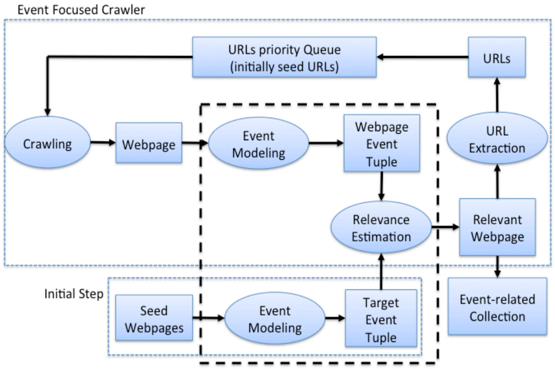

Figure 2. Event Focused Crawler (EFC) Architecture

EFC is modeling an event as **What** happened, **where**, and **when.** EFC uses information retrieval techniques (Vector Space/Probabilistic) to help find the **What** part and uses natural language processing techniques (Named Entity Recognition) to help find the **Where** and **When** parts.

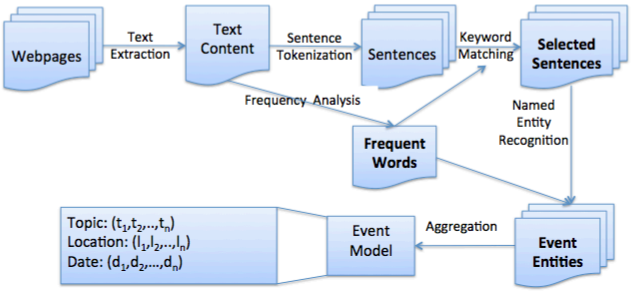

Figure 3. Steps for building event model from web collection

5. Conclusion

This paper explains semi-automatically collecting, curating, and archiving webpage collections, leveraging methods for event modeling and focused crawling. The event modeling covers especially identifying and representing events considering their What, Where, and When aspects.

Our work on focused crawling could be of benefit for Web archiving by:

1. Helping prepare lists of URLs to be archived (i.e., a focused crawler recommending a seed list)
2. Helping extend a collection automatically (using existing collections for machine learning type training of a focused crawler to find similar new webpages)
3. Analyzing and summarizing the produced event collections by using the developed event model

6. Acknowledgements

This material is based upon work supported by the US National Science Foundation under Grant No. IIS-1319578.

7. References

[1] Mohamed M. G. Farag and Edward A. Fox. Events Archive. Website for the IDEAL, CTRnet, and VT-DL-416 projects. http://www.eventsarchive.org. 2015

[2] Mohamed M. G. Farag and Edward A. Fox. Intelligent Event Focused Crawling. Proceedings of the 11th International ISCRAM Conference, University Park, Pennsylvania, USA, May 18-21, 2014t