Generating Stories From Archived Collections

Yasmin AlNoamany
University of California, Berkeley
Berkeley, California 94720, USA
yasminal@berkeley.edu

Michele C. Weigle and Michael L. Nelson
Old Dominion University
Norfolk, Virginia 23529, USA
{mweigle,mln}@cs.odu.edu

ABSTRACT

With the extensive growth of the Web, multiple Web archiving initiatives have been started to archive different aspects of the Web. Services such as Archive-It exist to allow institutions to develop, curate, and preserve collections of Web resources. Understanding the contents and boundaries of these archived collections is a challenge, resulting in the paradox of the larger the collection, the harder it is to understand. Meanwhile, as the sheer volume of data grows on the Web, "storytelling" is becoming a popular technique in social media for selecting Web resources to support a particular narrative or "story".

We address the problem of understanding archived collections by proposing the Dark and Stormy Archive (DSA) framework, in which we integrate "storytelling" social media and Web archives. In the DSA framework, we identify, evaluate, and select candidate Web pages from archived collections that summarize the holdings of these collections, arrange them in chronological order, and then visualize these pages using tools that users already are familiar with, such as Storify. Inspired by the Turing Test, we evaluate the stories automatically generated by the DSA framework against a ground truth dataset of hand-crafted stories, generated by expert archivists from Archive-It collections. Using Amazon's Mechanical Turk, we found that the stories automatically generated by DSA are indistinguishable from those created by human subject domain experts, while at the same time both kinds of stories (automatic and human) are easily distinguished from randomly generated stories.

CCS CONCEPTS

•**Information systems →Information retrieval;**

KEYWORDS

Web Archiving, Storytelling, Information Retrieval, Document Similarity, Archived Collections, Web Content mining, Internet Archive

ACM Reference format:
Yasmin AlNoamany and Michele C. Weigle and Michael L. Nelson. 2017. Generating Stories From Archived Collections. In *Proceedings of WebSci '17, Troy, NY, USA, June 25-28, 2017,* 10 pages.
DOI: http://dx.doi.org/10.1145/3091478.3091508

WebSci '17, Troy, NY, USA

DOI: http://dx.doi.org/10.1145/3091478.3091508

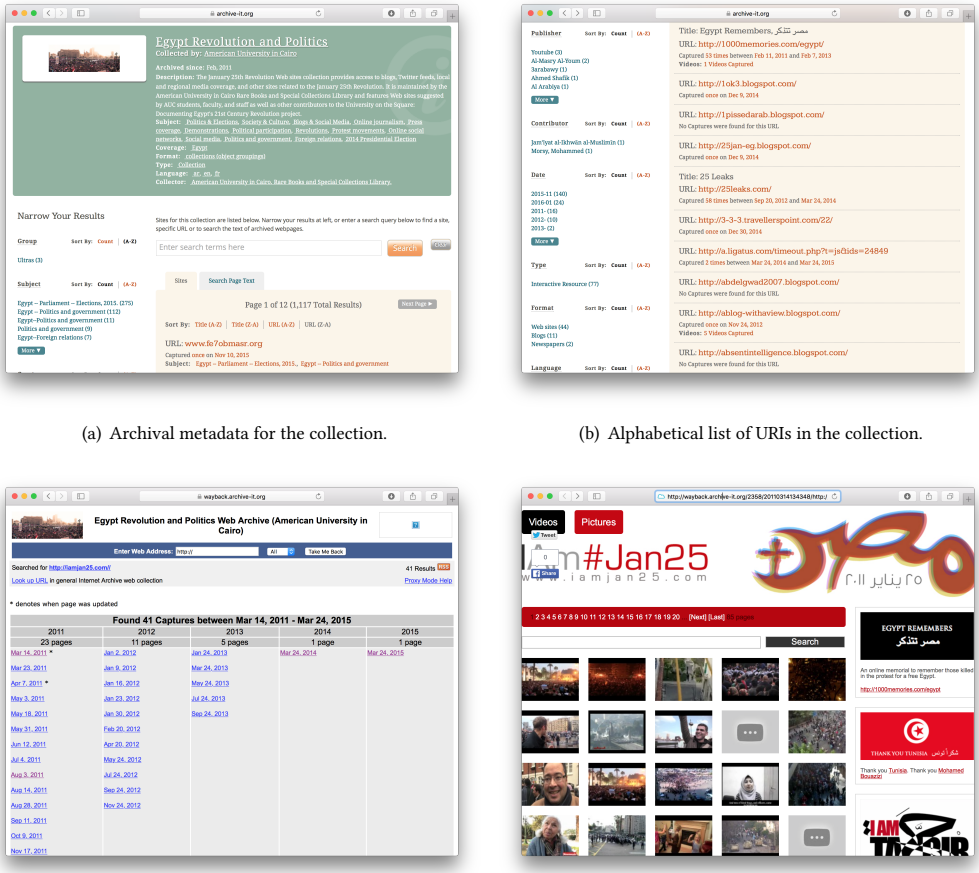
1 INTRODUCTION

Today's ordinary information will be tomorrow's resources for historical research. The content captured and published on the Web narrating the incidents and giving unfiltered insights for future generations and historians is important to clarify the exact turning points in history. Therefore, archiving Web pages into themed collections is an important method for ensuring these resources are available for posterity. Many initiatives exist to allow users to perform this task [27]. Many initiatives exist to allow people to archive Web resources into themed collections for ensuring these resources are available for posterity [6]. For example, Archive-It[1], a subscription service from the Internet Archive (IA)[2], allows institutions to develop, curate, and preserve topic-oriented collections of Web resources by specifying a set of seeds, Uniform Resource Identifiers (URIs), that should be crawled periodically. Archive-It provides users a listing of all seeds in the collection along with the number of times and dates over which each page was archived, as well as a full-text search of archived pages.

An archived collection can include hundreds of seed URIs. Over time, each of these URIs can be crawled hundreds or thousands of times, resulting in a collection having thousands to millions of archived Web pages. Understanding the contents and boundaries of a collection can be difficult [9], resulting in the paradox of the larger the collection, the harder it is to use. For example, a user of Archive-It interested in understanding the key events of the Jan. 25 Egypt Revolution will find multiple collections about this topic, and each of these collections may have a different focus. Aside from the brief metadata about the collection (Figure 1(a)), the interface mainly consists of a list of seed URIs in alphabetical order (Figure 1(b)), and for each of these URIs a list of the times when the page was archived (Figure 1(c)). It is not feasible for a user to figure out what is inside the collection without going through all the URIs in the collection and their relative archived copies. Understanding the essence of the collection from the current interface of Archive-It is not easy.

Providing a summary of the content of archived collections is a challenge because there are two dimensions that should be summarized: the URIs that comprise the collection (e.g., cnn.com) and the archived copies (called "mementos") of those URIs at different times (e.g., cnn.com@t_1, cnn.com@t_2,.., cnn.com@t_n). Either dimension by itself is difficult, but combined they present a number of challenges, and are hard to adapt to most conventional visualization techniques.

[1] http://www.archive-it.org/
[2] http://archive.org/



(a) Archival metadata for the collection.

(b) Alphabetical list of URIs in the collection.

(c) Archived copies of a URI in the collection.

(d) A copy of "Iam25Jan"

Figure 1: Current browsing and searching services for the "Egypt Revolution and Politics" collection in Archive-It.

We developed the Dark and Stormy Archives[3] (DSA) framework [2], which automatically extracts summary stories[4] from Archive-It collections to help the user to understand the collections. Events in these stories are summarized by sampling Web pages from the Archive-It collections, arranged in a narrative structure ordered by time, and replayed through storytelling social media interfaces such as Storify. By studying existing human-generated stories in Storify [3], we were able to profile different kinds of stories by examining the typical length (in terms of the number of resources included), time frames covered, structural metadata (e.g., page rank, images and video, social media vs. news) and other features. We used the structural characteristics of human-generated stories, with particular emphasis on "popular" stories (i.e., the top 25% of views, normalized by time available on the Web), that are applicable to the resources in Archive-It collections. For example, we generate stories automatically from archived collections with a typical length close to 28 (more or less based on the collection size).

What makes a good story is a matter of human judgment and is difficult to evaluate. We consider a story to be "good" if a person considers it to be indistinguishable from a human-generated story. Inspired by the Turing Test [26], we used ground truth dataset of hand-crafted stories from Archive-It collections and let humans select between the human-generated stories and the automatically generated stories. We consider our method to be a success if humans are as likely to choose the automatically generated story as they do the human-generated story. From this composite, we asked expert archivists to generate hand-crafted stories from Archive-It collection, then used Amazon's Mechanical Turk[5] to evaluate the automatically generated stories against the stories that were created by experts. Based on 332 comparisons by 30 unique Mechanical Turk workers (or "turkers") between human-generated stories and automatic stories, the results showed that at confidence level 95%, , turkers could not distinguish between the human-generated stories and the automatically generated stories ($p > 0.5$).

[3]Inspired by "It was a dark and stormy night", a well-known storytelling trope: https://en.wikipedia.org/wiki/It_was_a_dark_and_stormy_night/

[4]We use "story" in its current, loose context of social media, which is sometimes missing elements from the more formal literary tradition of dramatic structure, morality, humor, improvisation, etc.

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

2 RELATED WORK

Since the digitization process has started, most institutions, e.g., libraries and archives, have focused on storing digital collections and making them accessible online [11]. Most of the current digital collection interfaces are text-based search with very limited browsing features. Much research has been dedicated to developing visualizations for viewing and querying documents, and towards graphical browsing of the results [1, 15, 16, 28]. While Web archives are solutions for preserving the Web, they lack tools that allow users to understand the archived collections.

Our initial attempt to browse Archive-It collections and highlight the collections' underlying characteristics was applying four alternate visualizations (Bubble chart, Image plot with histogram, Timeline, Wordle) for the Archive-It interface [23]. The results are sufficient for those already with an understanding of what is in the collection, but they do not facilitate an understanding to those who are unfamiliar with collection.

Karmer-Smyth [19] developed ArchivesZ, an information visualization for archived collections inspired by the availability of structured data in the Encoded Archival Description [9] standard for encoding finding aids. The ArchivesZ prototype interface helps users explore the metadata that describes archival collections through searching for content by year and subject in a tightly coupled dual histogram interface. ArchivesZ gives users a visual representation of the total amount of content available in an archive on a given topic. It also visualizes the overlapping assignment of subjects terms to archival collections.

The UK Web Archive[6] provides a visualization for the collections through a 3D wall of sites allowing interaction through zooming.

One problem with the above approaches is that there is often an implicit assumption that everything in a collection is equally valuable and should be visualized. Some of the Web pages change frequently, some are near-duplicates, and some go off-topic and no longer contribute to the collection. Visualization techniques with an emphasis on recall (i.e., "here's everything in the collection") do not scale. Instead, we are informed by emerging trends in social media storytelling, which focus on a small number of exemplary pages (i.e., high precision) as chosen by a human, to sample from the collection by choosing representative pages that best exemplify the topic of the collection. Our work in selecting candidate Web pages leverages previous work in image collection summarization and video abstraction. Many image collection summarization techniques [5, 9, 10] divide the image collection by time, then cluster the images by content, and finally select a representative image from each cluster. In our framework, we take a similar approach to selecting representative mementos. Some video abstraction techniques [17, 21, 29] select keyframes that differ from each other in terms of their features, such as color, shape, motion, etc. In our work, we use text similarity to eliminate near duplicate mementos.

3 TYPES OF STORIES GENERATED FROM ARCHIVED COLLECTIONS

In the DSA framework, we apply IR and machine learning techniques to identify and select different sets of k mementos that compose stories, in which each story (S) provides an overview about the collection. So, we extract stories from a collection, $C \rightarrow S$, where $C \subset S$.

An archived collection has two dimensions. As we mentioned before, the collection is composed of a set of seed URIs and each seed has many copies through time. There may be multiple stories that convey different perspectives of the collection. In Table 1, we list four possible kinds of stories and name each story according to the change that happens to the URI and time:

- Fixed Page, Fixed Time (FPFT) is a different representation for the same Web site because of GeoIP, mobile, and other environmental factors [18]. It is generated using the same URI at a specific point of time with differences in the representation.
- Sliding Page, Sliding Time (SPST) is the broadest possible coverage of a collection. It is generated using different URIs at different times.
- Fixed Page, Sliding Time (FPST) is the evolution of a single page (or domain) through time. The possible scenario of this story is when a user wants to see how the story evolved over time from a specific Web site, e.g., cnn.com.
- Sliding Page, Fixed Time (SPFT) is different perspectives at a point in time. It is generated using different original URIs at nearly the same datetime.

Note that the FPFT story can not be supported by the current capabilities of Web archives because currently they do not provide users the ability to navigate representations by their environmental influences [18].

Table 1: Four basic story types (others may be possible).

		Time: fixed	Time: sliding
URIs:	fixed	differences in GeoIP, mobile, etc.	evolution of a single page (or domain) through time
	sliding	different perspectives at a point in time	broadest possible coverage of a collection

It is also possible that there are additional types of stories beyond those in Table 1, and we plan to investigate this in future work.

4 THE DARK AND STORMY ARCHIVES (DSA) FRAMEWORK

In this section, we present the Dark and Stormy Archives (DSA) Framework to select k archived pages that comprise a "story" that summarizes an Archive-It collection, arrange them in a narrative structure ordered by time (or any other type of story), then import them into existing storytelling tools or other visualizations.

4.1 Establish a Baseline

To support automatic story creation, we needed to better understand as a baseline the structural characteristics of human-generated stories. In our previous work [3], we investigated the structural characteristics of human-generated stories on Storify, with particular emphasis on "popular" stories. Upon analyzing 14,568 stories comprising 1,251,160 elements, we modeled the structural characteristics of the popular stories. We found that the popular stories

[6] https://www.webarchive.org.uk/ukwa/

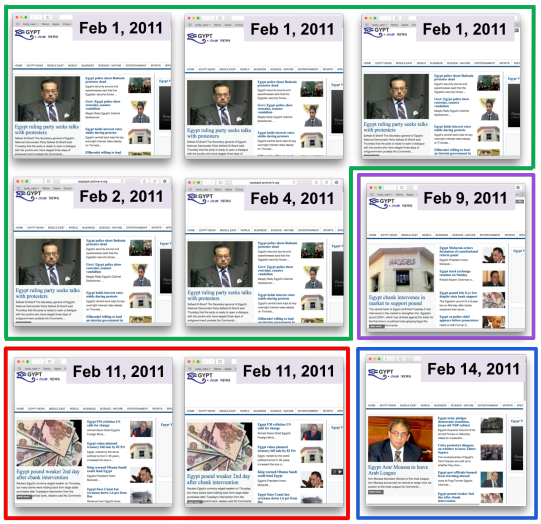


Figure 2: Snapshots of an Egyptian news Web site (http://news.egypt.com/en/) from the "Egypt Revolution and Politics" collection in Archive-It. Each group of similar mementos are grouped and annotated with the same color.

have a median value of 28 elements. This informs our framework for generating stories from archived collections that will be composed of a number of resources that is close to 28.

4.2 Reduce the Candidate Pool

Archive-It provides their partners with tools that allow them to build themed collections of archived Web pages hosted on Archive-It's machines. This is done by the user manually specifying a set of seed URIs that should be crawled periodically based on a predefined frequency set by the collection curator. This frequency may be daily, weekly, or even yearly. Due to the nature of Web evolution, some of these snapshots may change little or not at all. Some of the pages go off-topic and some other pages just become duplicates to other pages. We define off-topic pages as the Web pages that have changed through time to move away from the initial scope of the page. Currently, there are no content-based tools that allow curators to detect when seed URIs are off-topic. We apply the following steps on an archived collection to reduce the candidate pool of mementos:

(1) Exclude the off-topic pages from the collection.
(2) Exclude the (near-)duplicate mementos of each TimeMap, a list of mementos.
(3) Exclude the non-English language mementos.

In a previous work [4], we investigated and evaluated different approaches for detecting off-topic pages in individual TimeMaps on multiple Archive-It collections. In the DSA framework, we adopted the best performing method on Archive-It collections to eliminate the off-topic pages.

After excluding the off-topic pages, we eliminate (near-)duplicates. An example of duplicates in a TimeMap is illustrated in Figure 2. We select the first memento of the TimeMap and compare it to other subsequent mementos using Hamming Distance d. If the most recent memento exceeds a specific threshold α, which was determined empirically, it is selected to be the current memento that we compared to the subsequent mementos. We used 64-bit SimHash fingerprints with $k = 4$ to calculate the (near-)duplicates between Web pages in individual TimeMaps because of its time efficiency [14]. The goal is to generate a reduced TimeMap that contains only unique mementos of the URI.

Finally, we selected the English language mementos and excluded other languages. We detected the language of the content using the language detection library created by Shuyo [24] with precision $\geq 99\%$ [7, 24]. The DSA framework can be applied on pages with other languages, but currently, we evaluate English language pages only.

4.3 Select Good Representative Pages for Each Story

The previous step produces a set of reduced TimeMaps that have unique, relevant mementos to the topic of the collection. The following step is to evaluate and select the "best" representative k mementos, where k is much smaller than the number of mementos in the collection. As mentioned earlier, suggested values of k are determined by the results of previous work [3], and other tunable parameters will include the timeline of the desired story (which may exclude some portions of the collection), the percentage of damage of the memento (incomplete pages are not desirable candidates), the story type (cf. Table 1), etc. We combine all of mementos of all the TimeMaps into one set (the filtered mementos from all of the seeds) and then the following steps to select representative mementos for the story:

(1) Slice the collection dynamically and distribute the mementos equally on the slices.
(2) Cluster the pages in each slice.
(3) Evaluate and select the best representative page from each cluster based on multiple quality metrics.
(4) Put the selected pages in chronological order.
(5) Extract the metadata of the selected pages.
(6) Visualize the pages by leveraging storytelling tools, such as Storify.

We started by slicing the collection into a predefined number of slices Sc that is specified based on the number of mementos N in the collection after excluding the off-topic pages, non-English language pages, and the (near-)duplicates [2], so that:

If $|N| > 28$

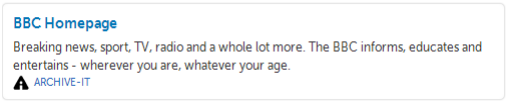
$$Sc = \lceil 28 + log_{10}|N| \rceil \quad (1)$$

Else

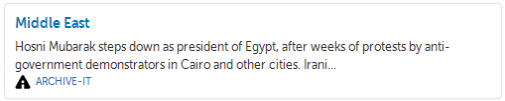
$$Sc = |N| \quad (2)$$

We then distribute the mementos equally on the slices and then cluster the mementos in each slice using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [12] based on their textual contents. DBSCAN does not require the specification of the number of clusters a priori, as opposed to k-means clustering [13]. The output of this step is a set of Cs clusters, where $Cs \geq Sc$.

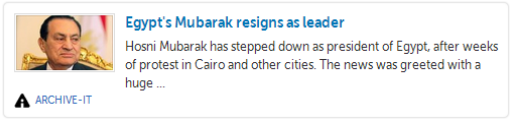
From each of the resulting clusters, we evaluate and select the best representative page based on multiple quality metrics. We specified the memento quality based on the amount of damage for



(a) Feb. 11, 2011: a memento of the homepage of BBC on Storify



(b) Feb. 11, 2011: a memento of the homepage of BBC Middle East section on Storify



(c) Feb. 11, 2011: a memento of the BBC article page on Storify

Figure 3: Storify creates better snippets from a specific article (i.e., deep links) than a homepage about the same event.

the memento and if the memento generates a visually attractive link preview when inserting into a tool like Storify. We adopted Brunelle's algorithm for assessing memento damage [8]. The quality of the visual link preview will tremendously affect the quality of the created story. When a user posts a link on social media networks, e.g., Facebook and Storify, a visual *snippet* with a title, a summary of the content, and an image is extracted from that link. These visual snippets are created from the HTML tags of the Web page. Based on experimenting the the generation of visual snippets for many different kinds of URIs [2], we discovered that social media sites can generate better snippets from articles that focus on only one topic (these articles also often have a long URI path length, e.g., cnn.com/a/b/c/2011/4/2), while they do not extract nice snippets from homepages that have an overview of multiple topics (these pages often have a short URI path length, e.g., cnn.com), as illustrated in Figure 3. Furthermore, the page category may affect the quality of the extracted snippets. For example, there are different kinds of URIs in which the extraction fails to capture information related to the topic of the collection such as URIs for pages on Facebook, Facebook accounts, Twitter accounts, Google groups, etc. When these pages are posted on Storify, the text of the snippet is extracted from the description of the profiles or pages.

Therefore, for specifying the quality of the memento, we weight each memento with quality measure M_q, which calculated as follows:

$$M_q = (1 - w_d \times D_m) + w_l \times M_l + w_c \times M_c \quad (3)$$

where D_m is the value of memento damage, M_l is URI level, and M_c is the URI category. We set level weight ($w_l = 0.45$), memento damage weight ($w_d = 0.40$), and category weight ($w_c = 0.15$). Setting these weights needs further testing with multiple collections. In the DSA framework, the value of M_l is normalized in the range of [0, 1]. For example, the M_l of cnn.com/a/b/c/2011/4/2 will be assigned 0.6 and $M_l = 0.1$ for cnn.com/. For calculating M_c, we adopted our previously proposed heuristic-based categorization [23], which classifies the URI based on its domain component, then assigns each category a weight $0 \leq M_c \leq 1$ based on how the category affects the snippet quality [2]. We give higher weights to news Web sites, video, social media posts, then blogs come next, and the lowest weight goes to Facebook pages, Twitter accounts, Google groups, etc.

After specifying the the best representative pages, extract the publish date of the page using the "Newspaper: Article scraping and curation" Python library [22]. It applies multiple strategies such as extracting the date from a URI or from the Web page metadata. If neither of these strategies succeed to estimate publishing date, we use the Memento-Datetime (the datetime the resource was crawled).

Finally, we order the mementos chronologically based on their dates and visualize the pages by leveraging storytelling tools. In our implementation, we used Storify, a popular platform for storytelling, to visualize the set of $k \approx 28$ mementos that represent the extracted story from the collection. Storify provides an API[7] that allows users to create and publish stories by sending objects of the elements of the stories in JSON format. Once a story is created and pushed to Storify, it can be edited and shared. For each story, we generate a JSON object that contains the metadata of the story, such as the story name and description, and the details of each element such as the hyperlink, the extracted title, etc. We override the favicon of the resource that is created by Storify because Storify uses the Archive-It favicon for all the pages regardless of the original source (see Figure 3).

5 EVALUATING THE DSA FRAMEWORK

In this section, we evaluate the automatically generated stories from archived collections.

5.1 Hand-crafted Stories from Archived Collections

We group Archive-It's collections into three main categories [4]. First, there are collections that are devoted to archiving governmental pages (e.g., all Web pages published by the State of South Dakota[8]). Second, there are collections that are event-based (e.g., Occupy Movement collection[9]). Third, there are theme-based collections (e.g., the Columbia Human Rights collection[10]).

We tested the DSA framework against event-based collections. We asked expert archivists, with the help of the Archive-It team and Archive-It partners, to generate hand-crafted stories from Archive-It collections. We provided them with guideline documents that contained instructions for generating stories from Archive-It collections by selecting 28 representative mementos (more or less based on the collection size) that best represent each collection. We showed them the type of stories that can be generated. We also provided them the criteria for selecting the mementos. They suggested 10 different collections to generate stories from (see Table 2).

The following is the list of the guidelines that we provided to the expert archivists for generating the stories:

[7] http://dev.storify.com/api/
[8] https://archive-it.org/collections/192/
[9] https://archive-it.org/collections/2950/
[10] https://archive-it.org/collections/1068/

Table 2: The characteristics of the collections used for the evaluation.

Collection	ID	Timespan	URIs	Mementos
2013 Boston Marathon Bombing	3649	2013/04/19 - 2015/03/03	318	1,907
Occupy Movement 2011/2012	2950	2011/12/03 - 2012/10/09	955	30,581
Egypt Revolution and Politics	2358	2011/02/01 - 2013/04/18	1,112	42,740
April 16 Archive	694	2007/05/23 - 2008/04/28	88	362
2013 Government Shutdown	3936	2013/10/22 - 2013/10/22	186	246
Russia Plane Crash Sept 2011	2823	2011/09/08 - 2011/09/15	104	558
Wikileaks 2010 Document Release Collection	2017	2010/07/27 - 2013/08/26	41	1,126
Earthquake in Haiti	1784	2010/01/20 - 2011/02/27	132	967
Brazilian School Shooting	2535	2011/04/09 - 2011/04/14	650	1,492
Global Health Events	4887	2014/10/01 - 2015/12/21	169	3,026

Table 3: The number of resources in the 23 stories (10 SPST, 6 SPFT, 7 FPST) generated by domain experts and from the DSA framework.

		SPST		SPFT		FPST	
Collection	ID	Human	Automatic	Human	Automatic	Human	Automatic
2013 Boston Marathon Bombing	3649	28	29	28	25	7	5
Occupy Movement 2011/2012	2950	16	45	9	20	9	7
Egypt Revolution and Politics	2358	16	20	11	17	12	7
April 16 Archive	694	17	32	14	19	5	4
2013 Government Shutdown	3936	17	27	14	15	-	-
Russia Plane Crash Sept 2011	2823	28	25	27	23	-	-
Wikileaks 2010 Document Release Collection	2017	25	32	-	-	7	10
Earthquake in Haiti	1784	28	34	-	-	11	14
Brazilian School Shooting	2535	26	24	-	-	23	20
Global Health Events	4887	36	34	-	-	-	-

- The representative mementos should be selected from within the collection. There should not be any memento from outside the collection.
- The default value for the number of selected mementos is $k \approx 28$. This value can be more or less based on the nature and size of each collection.
- We expect to have three generated stories out of each collection. Depending on the nature of the collection, some kind of stories may not be applicable. For those collections, please specify if any of the previous kinds of stories cannot be created.
- You can choose a specific time period for generating the story. If the collection spans many years, you can choose a subset of the timespan of the collection.

We also put criteria for selecting the mementos: the language of the memento should be in English; the memento should be on-topic (the content is related to the topic of the collection); the memento should produce a visually attractive snippet on Storify, an article (cnn.com/a/b/12/2015) is more preferred than a homepage (cnn.com); the memento should not be a (near-)duplicate of another memento in the list; a memento with no missing resources is a better choice than a memento that is missing resources.

Along with the criteria of the stories and the selected mementos within each story, we illustrated to the Archive-It team the suggested possible types of stories that can be generated from each collection.

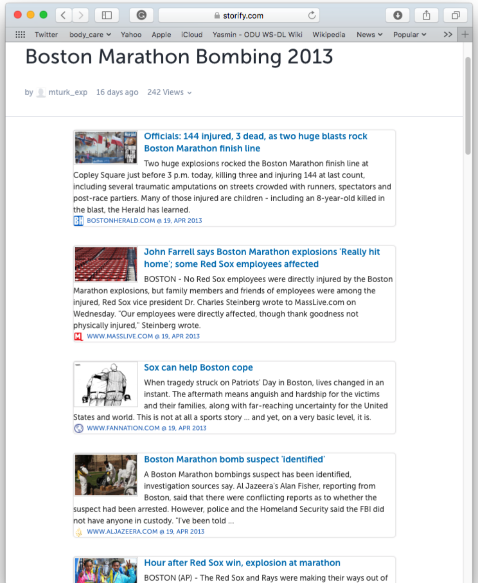
The domain experts provided us with lists of mementos for 23 different stories from the 10 different collections (see Table 3). Table 3 also shows the number of resources per story that were generated by experts and by the DSA framework. An example of a manually generated story by archivists from the Boston Marathon Bombing collection is shown in Figure 4(a).

There were some collections that spanned a short period of time, so the archivists did not provide the FPST stories for these collections (for example, the "Brazilian School Shooting", which spans over three days only). Another reason for not generating the FPST story is that none of the seeds of the collection change over time (e.g., news articles). For example, the seed URIs of "Russia Plane Crash Sept 2011" collection are all news articles which do not evolve over time.

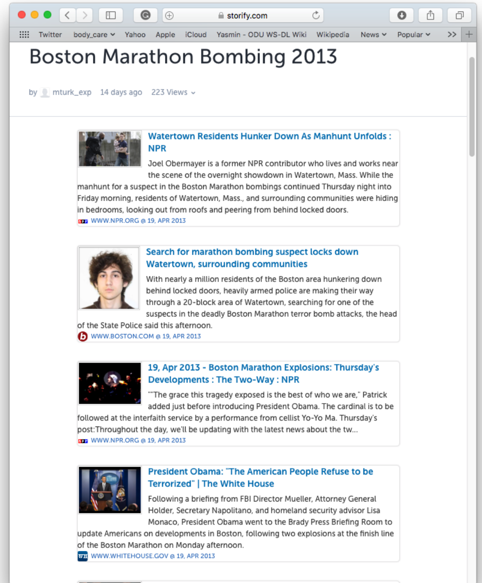
5.2 Automatically Generated Stories from Archived Collections

We then applied the steps of the DSA framework (Section 4) on the set of suggested collections in Table 2. We automatically generated 23 stories[15] from the collections (see Table 3). The FPST stories and the SPFT stories require input parameters such as the TimeMap for FPST stories and time frame for SPFT stories. In these stories, we

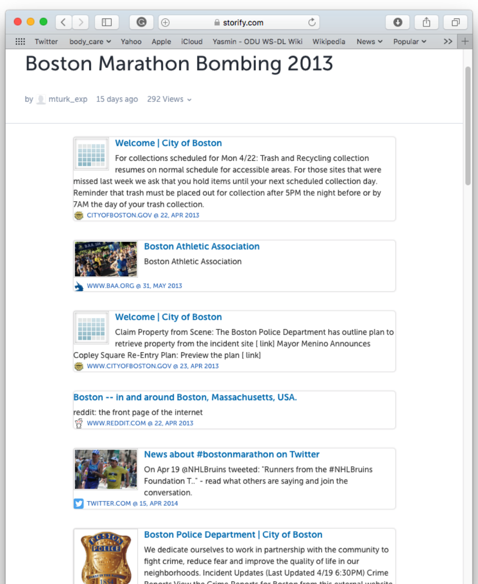
[14]https://storify.com/mturk_exp/3649b1s-57218803f5db94d11030f90b
[14]https://storify.com/mturk_exp/3649b0s
[14]https://storify.com/mturk_exp/3649b2s-57227227bb79048c2d0388dc
[14]https://storify.com/mturk_exp/3649bads
[15]Links to these stories are available at https://github.com/yasmina85/DSA-stories



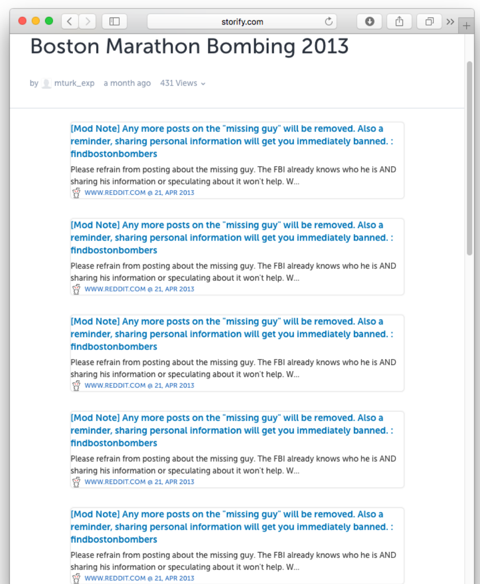
(a) Human-generated story[11].



(b) Automatically-generated story[12].



(c) Randomly-generated story[13].



(d) Poorly-generated story[14].

Figure 4: Example for SPST stories from the Boston Marathon Bombing collection.

use the same parameters that were used in the human-generated stories and input them to the DSA (Table 2). The SPST stories do not require any parameters because they represent a broad summary for the whole collection from all the seed URIs at different times. An example of an automatically generated story by the DSA framework is illustrated in Figure 4(b).

The number of the resources in the generated stories are presented in Table 3. Note that although the Egypt Revolution and

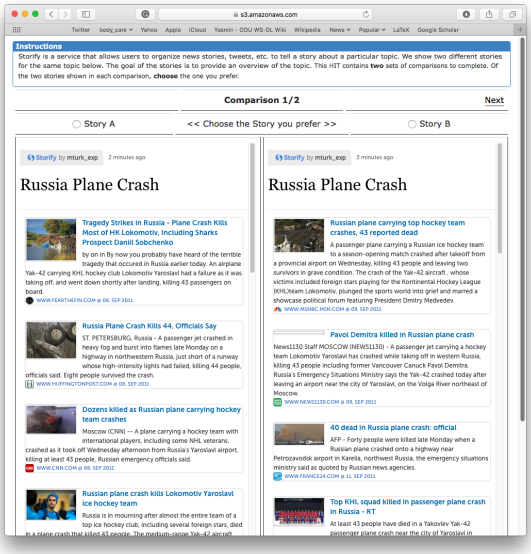


Figure 5: A sample HIT that shows two stories that turkers evaluate and select their preferred story. Each HIT contains two comparisons.

Politics collection is the largest collection in the dataset, the resulting number of the resources for the SPST story from this collection is just 20 mementos. That is because we selected the pages from within the same time frame (2011/02/01-2011/02/14) that was used for the human-generated story.

5.3 Random and Poor Stories

We use randomly generated stories to be compared against the human-generated stories and the automatically generated stories as a baseline. In other words, we expect that both the automatically generated stories and human generated stories will perform better than random stories. We selected $k \approx 28$ mementos randomly (see Figure 4(c)) from the set of mementos in each collection as a baseline for evaluating the automatically generated stories. The selection was done on the mementos in the collection before excluding the off-topic or the duplicates. The selected mementos were not sorted chronologically in the generated stories.

We generated poor stories by randomly selecting a memento from collection's TimeMap and repeating this memento 28 times. This story represents a control to ensure that the turkers do not choose randomly between the stories.

We used the same extraction methods for visualizing the human-generated stories, automatically generated stories, randomly generated stories, and poorly generated stories on Storify.

5.4 Experiment Setup

We use Mechanical Turk to compare four types of stories (human-generated, automatically generated, randomly generated, poorly generated), asking turkers to choose between two stories at a time.

Our goal is to assess if the automatically generated stories by the DSA framework are indistinguishable from the human-generated stories. We provided turkers a description of a simple task to perform (a Human Intelligence Task, or HIT), choosing their preferred story (see Figure 5). We provided a simple generic description for the task as follows:

> *Storify is a service that allows users to organize news stories, tweets, etc. to tell a story about a particular topic. We show two different stories for the same topic below. The goal of the stories is to provide an overview of the topic. This HIT contains two sets of comparisons to complete. Of the two stories shown in each comparison,* ***choose*** *the one you prefer.*

Each HIT consists of two comparisons, in which one of the two comparisons was a control, a comparison between one of the stories and a poorly generated story. We reject the HITs where users selected a poorly generated story (i.e., a false positive selection).

To reduce the cognitive load of the task, we assigned one comparison for each HIT along with the comparison that includes the poor story. Therefore, for evaluating one story, we have three HITs as follows:

HIT_1 : human vs. automatic, human vs. poor

HIT_2 : human vs. random, human vs. poor

HIT_3 : random vs. automatic, automatic vs. poor

We ensured that the position of each pair of composites was reversed among different stories to ensure there was not a bias in the HIT layout. We posted 69 HITs to evaluate 23 different stories. For each HIT, we required 15 turkers with "master" qualification requirements[16]. Based on many studies for deciding the number of participants in user studies, group sizes between eight and 25 are typically good numbers for conducting comparative studies [20, 25]. We chose to use 15 participants for each HIT in our experiment. We rejected the HITs in which the submissions contained poorly generated stories and the HITs that were completed in less than 10 seconds. TWe rejected a total of 46 HITs. In total, we had 989 out of 1,035 (69×15) valid HITs. These HITs were performed by 30 unique Master level turkers. We awarded the turker $0.50 per HIT. The turkers took seven minutes on average to complete the selections of the two comparisons.

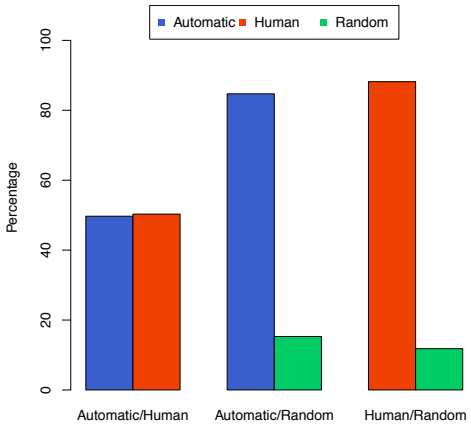
Table 4: The results of comparing human-generated stories versus automatically generated stories.

	Selections	Human	Automatic
SPST	142	50.7%	49.3%
SPFT	87	46.0%	54.0%
FPST	103	51.5%	48.5%

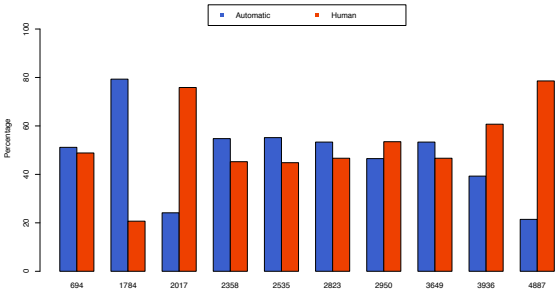
5.5 Results

Figure 6(a) shows a summary of the results of the turkers selections for the three comparisons: human vs. automatic, random vs. automatic, and human vs. random. The results in Figure 6(a) show that both the automatically generated stories and the human-generated

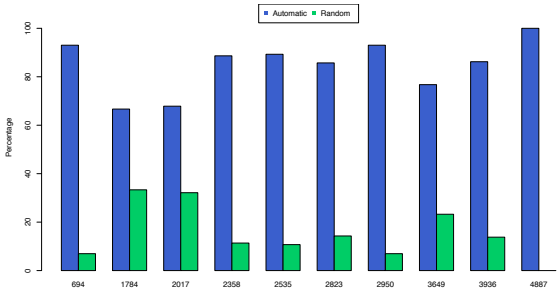
[16] https://www.mturk.com/mturk/help?helpPage=worker#what_is_master_worker



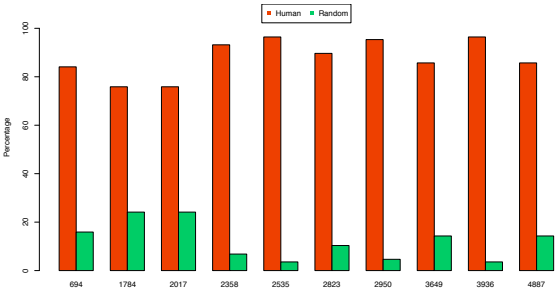
(a) A summary of the results.



(b) Automatic versus Human per collection.



(c) Automatic versus Random per collection.



(d) Random versus Human per collection.

Figure 6: DSA-generated stories are indistinguishable from human-generated stories, and both are distinguishable from random stories.

stories were selected $\approx$ 50% of the time. The figure also shows that the automatic stories are better than the randomly generated stories. Based on the results of the two-tailed t-test on the number of votes received, we found that at confidence level 95% the automatically generated stories with $mean = 7.17$ are indistinguishable from the human-generated stories with $mean = 7.26$ ($p = 0.9134$, $t = 0.1094$, $df = 43.9$). However, at confidence level 95%, the automatically generated stories with $mean = 12.04$ and the human-generated stories with $mean = 12.65$ are significantly different from the randomly-generated stories with $mean \approx 2$ ($p < 2.2e{-}16$).

We zoom in on the results of the human-generated sto ries versus the automatically generated stories to interpret the results based on the different types of stories (SPST, SPFT, FPST). Table 4 shows that for all types of stories, the percentages of the turkers preferences to human and automatic stories are close. We applied a two-sided paired t-test on the samples based on the story type. We found that at confidence level 95% there is no significant difference ($p >$ 0.5) between the human-generated stories and the automatically generated stories for all the types of the stories. However, the difference between the automatically generated stories and the randomly-generated story is statistically significant ($p < 0.001$) for all the types of stories at 95% confidence level. There is also a significance difference between the randomly generated stories and the human-generated stories ($p < 0.001$) at 95% confidence level

We show the results of the turkers' preferences for the three selections for each collection in Figure 6. Figure 6(b) shows that for most of the collections, the automatically generated stories are indistinguishable from the human-generated stories. There are two collections that human-generated stories were selected more than the automatically generated stories: the "Wikileaks 2010 Document Release" (2017) and "Global Health Events" (4887). The automatically generated stories for the "Earthquake in Haiti" (1784) were preferred by turkers. Further investigation with more

collections is required to test if the type of collections affects a human's selection.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we presented the DSA framework, in which we identify, evaluate, and select candidate mementos to support the events of the stories. Our goal is to allow users to get many perspectives about the collection and also about how the story of the collection has evolved over time. We leverage narrative visualizations and storytelling tools, such as Storify, to visualize the created stories and demonstrate how they have evolved over time. We evaluated the stories generated by the DSA framework. We obtained a ground truth dataset of 23 stories that were generated manually from 10 Archive-It collections by expert archivists. We used Amazon's Mechanical Turk to compare the automatically generated stories with the human-generated stories. Based on 332 comparisons by 30 unique turkers between human-generated stories and automatic stories, the results showed that at confidence level 95%, the automatically generated stories are indistinguishable from the human-generated stories ($p > 0.5$). We also created random stories as a baseline for the automatic stories. The results show that the turkers were able to distinguish the random stories from the automatic and the human stories ($p < 0.001$). The code and gold standard dataset are available at https://github.com/yasmina85/DSA-stories.

We provided preliminary evaluation for the stories generated by the DSA framework. Although the humans were not able to distinguish the automatically generated stories from the human-generated stories, future research should investigate the usefulness of the generated stories and evaluate the discovery tasks for people given the summarized stories. Furthermore, we plan to collaborate with humanities researchers to conduct user studies on important events, e.g., the Arab Spring, and check if a specific kind of story provides the best insight into the events and the corresponding collections. For example, how do the Sliding Page, Fixed Time stories help humanities researchers to get different perspectives about news coverage and how much time is saved from manual search by providing them this kind of story.

7 ACKNOWLEDGMENTS

This work supported in part by the Institute Museum and Library Services (LG-71-15-0077-15). We thank the Archive-It team and partners for creating the gold standard data set.

REFERENCES

[1] Christopher Ahlberg, Ben Shneiderman, and Ben Shneidennan. 1994. Visual information seeking: tight coupling of dynamic query filters with starfield displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '94)*. 313–317. DOI: https://doi.org/10.1145/191666.191775

[2] Yasmin AlNoamany. 2016. *Using Web Archives to Enrich the Live Web Experience Through Storytelling*. Dissertation. Old Dominion University.

[3] Yasmin AlNoamany, Michele C Weigle, and Michael L Nelson. 2016. Characteristics of Social Media Stories. What makes a good story? *International Journal on Digital Libraries* 17 (2016), 239–256. DOI: https://doi.org/10.1007/s00799-016-0185-3

[4] Yasmin AlNoamany, Michele C Weigle, and Michael L Nelson. 2016. Detecting Off-Topic Pages Within TimeMaps in Web Archives. *International Journal on Digital Libraries* 17 (2016), 203–221. DOI: https://doi.org/10.1007/s00799-016-0183-5

[5] Ahmed AlSum and Michael L Nelson. 2014. Thumbnail Summarization Techniques for Web Archives. In *Proceedings of the 36th European Conference on Information Retrieval (ECIR '14)*. 299–310. DOI: https://doi.org/10.1007/978-3-319-06028-6_25

[6] Jefferson Bailey, Abigail Grotke, Kristine Hanna, Cathy Hartman, and Nicholas Taylor. 2004. Web Archiving in the United States: A 2013 Survey. http://www.digitalpreservation.gov/documents/NDSA_USWebArchivingSurvey_2013.pdf. (2004).

[7] Ralf D. Brown. 2013. Selecting and Weighting N-Grams to Identify 1100 Languages. In *Text, Speech, and Dialogue (Lecture Notes in Computer Science)*, Vol. 8082. 475–483.

[8] Justin F. Brunelle, Mat Kelly, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson. 2015. Not All Mementos Are Created Equal: Measuring The Impact Of Missing Resources. *International Journal of Digital Libraries* 16, 3 (2015), 283–301. DOI: https://doi.org/10.1007/s00799-015-0150-6

[9] Michelle Chang, John J. Leggett, Richard Furuta, Andruid Kerne, J. Patrick Williams, Samuel A. Burns, and Randolph G. Bias. 2004. Collection Understanding. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '04)*. 334–342. DOI: https://doi.org/10.1145/996350.996426

[10] Wei-Ta Chu and Chia-Hung Lin. 2008. Automatic Selection of Representative Photo and Smart Thumbnailing Using Near-duplicate Detection. In *Proceedings of the 16th ACM International Conference on Multimedia (MM '08)*. ACM Press, 829–832. DOI: https://doi.org/10.1145/1459359.1459498

[11] Laura Deal. 2015. Visualizing Digital Collections. *Technical Services Quarterly* 32, 1 (2015), 14–34. DOI: https://doi.org/10.1080/07317131.2015.972871

[12] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, 226–231.

[13] John A Hartigan and Manchek A Wong. 1979. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 100–108.

[14] Monika Henzinger. 2006. Finding near-duplicate web pages: a large-scale evaluation of algorithms. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Press, 284–291. DOI: https://doi.org/10.1145/1148170.1148222

[15] Keyun Hu. 2014. *VisArchive: A Time and Relevance Based Visual Interface for Searching, Browsing, and Exploring Project Archives (with Timeline and Relevance Visualization)*. Dissertation. University of Victoria.

[16] Keyun Hu, Melanie Tory, Sheryl Staub-French, and Madhav Prasad Nepal. 2016. VisArchive: a time and relevance based visual interface for searching, browsing and exploring project archives. *Visualization in Engineering* 4 (2016). DOI: https://doi.org/10.1186/s40327-016-0036-8

[17] Hang-Bong Kang. 2002. *Video Abstraction Techniques for a Digital Library*. IGI Global. 120–132 pages.

[18] M Kelly, JF Brunelle, MC Weigle, and ML Nelson. 2013. A Method for Identifying Personalized Representations in Web Archives. *D-Lib Magazine* 19 (2013), 2. DOI: https://doi.org/10.1045/november2013-kelly

[19] J Kramer-Smyth, M Nishigaki, and T Anglade. 2007. ArchivesZ: Visualizing Archival Collections. http://archivesz.com/ArchivesZ.pdf. (2007).

[20] Ritch Macefield. 2009. How to specify the participant group size for usability studies: a practitioner's guide. *Journal of Usability Studies* 5, 1 (2009), 34–45.

[21] Jung Oh, Quan Wen, Sae Hwang, and Jeongkyu Lee. 2004. Video abstraction. *Video data management and information retrieval* (2004), 321–346.

[22] Lucas Ou-Yang. 2013. Newspaper: Article scraping & curation. http://newspaper.readthedocs.io/. (2013).

[23] Kalpesh Padia, Yasmin AlNoamany, and Michele C. Weigle. 2012. Visualizing Digital Collections at Archive-It. In *Proceeding of the 12th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '12)*. 437–438. DOI: https://doi.org/10.1145/2232817.2232821

[24] Nakatani Shuyo. 2012. Language Detection Library for Java. http://code.google.com/p/language-detection/. (2012).

[25] Janet M Six and Ritch Macefield. 2016. How to Determine the Right Number of Participants for Usability Studies. UXmatters, http://www.uxmatters.com/mt/archives/2016/01/how-to-determine-the-right-number-of-participants-for-usability-studies.php. (2016).

[26] Alan M Turing. 1950. Computing machinery and intelligence. *Mind* 59, 236 (1950), 433–460.

[27] Khoi Duy Vo, Tuan Tran, Tu Ngoc Nguyen, Xiaofei Zhu, and Wolfgang Nejdl. 2016. Can We Find Documents in Web Archives Without Knowing Their Contents?. In *Proceedings of the 8th ACM Conference on Web Science (WebSci '16)*. 173–182. DOI: https://doi.org/10.1145/2908131.2908165

[28] M Whitelaw. 2009. Exploring Archival Collections with Interactive Visualisation. http://www.eresearch.edu.au/docs/2009/era09_submission_74.pdf. In *Proceedings of E-Research Australasia Conference*.

[29] Hong Jiang Zhang, Jianhua Wu, Di Zhong, and Stephen W Smoliar. 1997. An integrated system for content-based video retrieval and browsing. *Pattern recognition* 30, 4 (1997), 643–658.