Media Fragments Indexing using Social Media

Yunjia Li[1], Raphaël Troncy[2], Mike Wald[1], and Gary Wills[1]

[1] University of Southampton, UK
yl2@ecs.soton.ac.uk, mw@ecs.soton.ac.uk, gbw@ecs.soton.ac.uk
[2] EURECOM, Sophia Antipolis, France,
raphael.troncy@eurecom.fr

Abstract. With more and more video resources shared on the Web, the practice of sharing a video object from a certain time point (deep-linking) has been implemented by many video sharing platforms. With so many media fragments created, annotated and shared, however, indexing video objects on a fine-grained level on the Web scale is still not implemented by major search engines. To solve this problem, this paper proposes Twitter Media Fragment Indexer, which monitors the Tweet text and uses the embedded URLs pointing to video fragments as the media to create index for media fragments. In this paper, we show a preliminary evaluation that thousands of media fragments can be successfully indexed using this system. We are planning to expand the indexer in a larger scale and prove that millions of media fragments can be indexed by major search engines in this way.

Keywords: media fragment, media annotation, seo, schema.org

1 Introduction

Search Engine Optimisation (SEO) for videos is obtaining more attention with the booming of online video sharing applications, such as YouTube[3] and Dailymotion[4]. Many of those applications have deep-linking to a certain temporal fragment of the video and some of them also provide the search function for media fragments within their applications based on the text resources (such as closed captioning) linked to them[5]. However, SEO for general search engines for multimedia resources in a fine-grained level is still far away from us. Theoretically, with the the help of the deep-linkings, search engines should be able to match the keywords that users input and return the media fragment URI as the search results. However, this is not the case for most search engines, such as Google and Bing[6]. By analysing how the videos and related annotations are presented on the landing page of those video sharing applications, Li *et al.* [4]

[3] http://www.youtube.com
[4] http://www.dailymotion.com
[5] One example is that YouTube allow users to search the closed captioning (cc) timely aligned with the video.
[6] http://www.bing.com

finds out that the issue lies in that the annotations about different media fragments are sharing the same landing page, so that they do not have their own URIs to be indexed by search engines. To solve this problem, Media Fragment Indexing Framework [4], which uses Google Ajax Crawler, is developed to prepare a snapshot page for each media fragment and let it indexed by Google (See Section 2.2 for detailed explanation for the framework).

Even though the evaluation has shown that the framework can successfully enable Google to index media fragments, it still relies on the video sharing applications to adopt this framework and ask users to create the media fragment annotations manually, which is not the case currently for major video sharing platforms. Thus, this framework needs to be extended by more media fragment annotations collected automatically from massive users. In this paper, we propose the using of social media, Twitter for example, to help the indexing of media fragments using the Media Fragment Indexing Framework. The basic approach is to treat the text of each Tweet, which contains media fragment URI from major video sharing applications, as the annotation of that media fragment, so that we can develop a programme to monitor such Tweets and submit them to Google for media fragment indexing.

In the rest of this paper, Section 2 will introduce the background knowledge including W3C Media Fragment URI 1.0 (basic) and the Media Fragment Indexing Framework. Section 3 conducts a survey on which video sharing platforms have the deep-linking function similar to YouTube, so that they could be monitored to obtain the media fragment URIs. Based on the survey results, Section 4 develops a demo application, named Twitter Media Fragment Indexer, to collect the Tweets with media fragment URIs and make the media fragment URIs indexed by Google via the text in Tweets using the Media Fragments Indexing Framework. Finally, Section 5 makes the conclusions and points out some future research directions.

2 Related Work

This section will introduce some previous work as the background knowledge. We will also discuss the issues of the current solution of media fragment indexing and point out why using social media can improve the indexing results.

2.1 W3C Media Fragment URI

The W3C Media Fragment Working Group in Video in the Web Activity[7] have collected a wide range of use cases of using media fragments and proposed Media Fragment URI 1.0 (basic) [6] (W3C-MFURI). Media Fragment URI 1.0 (basic) is a W3C recommendation, which supports the addressing of image, audio and video along two major dimensions: temporal and spatial. Two more dimensions,

[6] `http://goo.gl/dPc81`
[7] `http://www.w3.org/2008/WebVideo/`

track and id (such as chapter 1, section 3, etc) are further defined in a W3C working draft named Media Fragments 1.0 URI (advanced) [5]. The information about each dimension is encoded in URIs using hash fragments following a certain format.

Many online video sharing platforms have (partially-)implemented the W3C-MFURI. For example, YouTube has launched facilities to annotate parts (temporal and spatial) of a video clip. Users can right click on a playing YouTube video and "copy video url at current time". The copied URI will have a URI fragment starting with "t=XXs" and YouTube can play from that time point. In Nov, 2013, YouTube introduced a new feature which allows users to "tag" spatial area in a temporal point. Each tag that a user creates has a URL pointing to a landing page hosted via Clickberry[8] then this link can be shared via Facebook and Twitter. Here is an example of the tag URL:

`https://clickberry.tv/video/6dafe30e-dcb8-44b8-8190-32be8249a297`

With the development of embedding semantic markups techniques in (X)HTML, such as RDFa [1] and Microdata [2], the web pages with embedded semantic markups can obtain better ranking or get highlighted in the search results by "traditional" Web search engines. It is the same case for online video objects. For example, Schema.org[9] defines *VideoObject* as the primary object for embedding structured description in the web pages which mainly serve videos. Google also suggests the use of video sitemap[10] to highlight videos in the search results.

2.2 Introduction of Media Fragments Indexing Framework

This section will briefly introduce the Media Fragments Indexing Framework developed in [4]. Google has developed a framework to crawl Ajax applications (see Figure 1). If the "hashbang" token ("#!") is included in the original URL, Google crawler will know that this page contains Ajax content. Then the crawler will request the "ugly URL", which replaces the hashbang by a query parameter "_escaped_fragment_". On receiving this "ugly URL" request, the server can return the snapshot page representing the page after the dynamic information is fully generated by javascript. The content in the snapshot page will be indexed in for the original "pretty URL".

Figure 2 explains how this framework could be used to index media fragments. The returned page in step 4 only contains keywords related to fragment "t=3,7". In the Google index, the "pretty media fragment URLs" are associated with the snapshot page. So what Google actually indexed is the URL of the replay page with hashbang and W3C-MFURI syntax attached. Step 8 still returns the whole page, but in step 9, the fragment will be passed to the URL representing the real location or the service which delivers the multimedia file. For example,

[8] `http://clickberry.tv`
[9] `http://schema.org`
[10] `https://support.google.com/webmasters/answer/80472?hl=en`
[10] `http://goo.gl/dPc81`

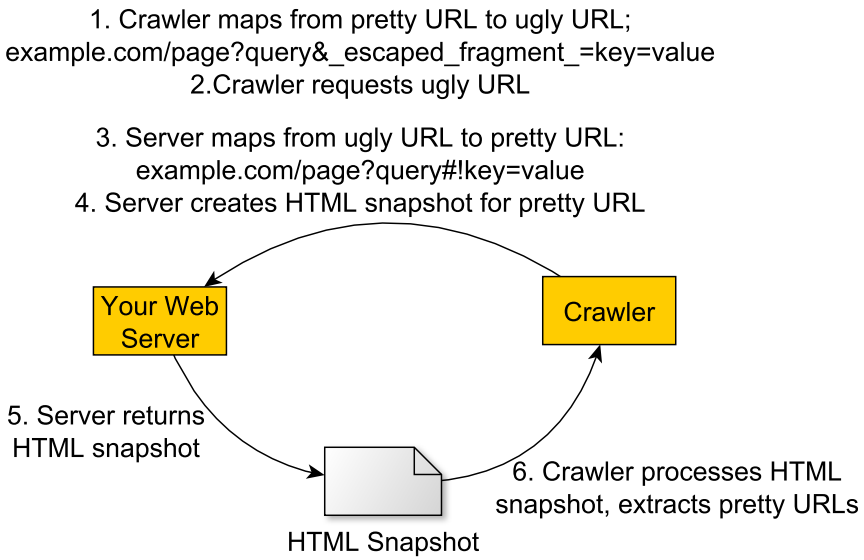

Fig. 1. Workflow of Google Ajax crawler

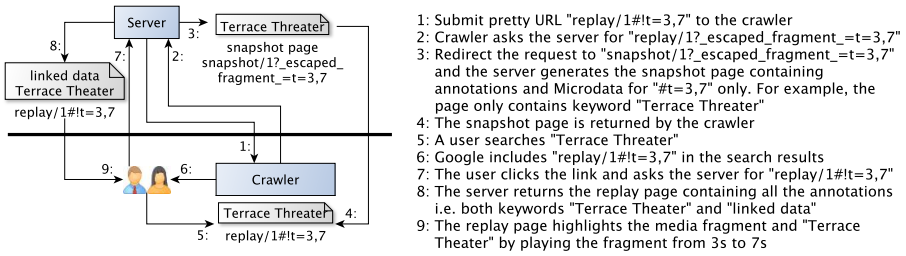

Fig. 2. Media Fragment Indexing Framework based on Google Ajax Crawler

if the request URL is *example/replay/1#!t=3,7*, the fragment "#t=3,7" will be attached at the back of *example2/1.ogv*, which is the video embedded in the replay page. Hashbang is not a valid syntax in W3C-MFURI specification, so developers need to parse the information in the hashbang URL before attaching the fragment to the the URL of the actual multimedia file. Then the embedded player needs to play the fragment from 3s to 7s controlled by javascript and the corresponding annotations are highlighted directly. In this case, step 6 will return the URL of the media fragment instead of the replay page. This design not only makes sure media fragments are indexed precisely with the keywords related to it, but also preserves the existing landing page.

2.3 Discussion

Even though the Ajax crawler is designed for crawling Ajax content on the Web scale, the scalability of the Media Fragment Indexing Framework introduced above will depend on how many users in the application can manually make annotations and link them to media fragments. In other words, unless the application has millions of users, the Media Fragment Indexing Framework can only index a limited number of videos.

It is a common practice now to share video on social media applications, such as Twitter. As some video sharing applications expose the deep-linkings into the video as URLs and allow users to share them on Twitter, it is a straight-forward thought that a Tweet message containing media fragment URI(s) can be treated as an annotation to the media fragment(s). Thus, if the Tweets can be monitored and filtered by whether media fragments are contained in message, those Tweets can be automatically uploaded as the input of the Media Fragment Indexing Framework. In this way, massive amount of videos can be indexed on the media fragment level. The following two sections will present the rationale of Twitter Media Fragment Indexer and show some preliminary evaluation of the system.

3 W3C-MFURI on Video Sharing Platforms

Theoretically, any URL shared on Twitter could be a media fragment URI. Ideally, we should monitor all the Tweets and check whether a media fragment URI is included in the message. However, this method is not realistic in that it is difficult to automatically decide whether a URL in the message is a media fragment URI. Some URLs may use the syntax similar to W3C-MFURI, but they have nothing to do with media fragments (false positive cases), for example, *http://www.example.org/1234#t=23*

Unless the HTML page is manually examined, it is hard to decide whether this URI is about a media fragment, but *#t=23* is indeed a valid W3C-MFURI syntax. We developed a programme using Twitter firehose API[11] to examine every Tweet message within one minute and the programme uses the Media Fragment URI Parser[12] recommended by the Media Fragment Working Group to decide whether a URL in the Tweet is a valid media fragment URI. We can see from Table 1 that only 2 out of 49 recognised "media fragment URIs" are real. Even though some URLs encode *t=xx* or *track=xx* as URI hash or query, they are obviously not media fragment URIs. This programme shows that, by parsing every URL shared on Twitter only, the false positive rate of the general monitor is too high. So the Twitter Media Fragment Indexer needs to apply other methodologies that are cost efficient and less error-prone.

Nowadays, most of the videos users watch online are hosted on major video sharing platforms. So it is reasonable to expect that most of the media fragments shared in Twitter are coming from those websites. If the indexer can filter the

[11] `https://dev.twitter.com/docs/api/1.1/get/statuses/firehose`
[12] `https://github.com/tomayac/Media-Fragments-URI/`

Table 1. Twitter Media Fragment Monitor Using Twitter Firehose API

Total Tweets examed	4356
Media Fragment URLs recognised by the parser	49
Real Media Fragments URLs	2
False positive rate	95.9%

Tweets containing URIs from those domains and parse the URIs according to the media fragment syntax defined in each website, the parsing results will be more accurate. Of course, some of the media fragments could be missing because the domains which host the videos are not monitored by the indexer. So, as the first step to use Twitter as the data source of media fragment indexing, a survey need to be conducted to find out which online video sharing platforms have actually implemented or partially implemented the notion of media fragment, so that their media fragments can be shared via Twitter. The rest of this section will design a survey for this purpose and analyses the survey results.

3.1 Methodology

The first step in the methodology is to decide which video sharing application(s) need to be investigated. There is a Wikipedia page "List of video hosting services"[13], where the major video hosting websites are listed. This experiment slightly modifies the list by adding a couple of well-known applications, such as TED.com[14] and videolectures.net[15], and removing the ones that are not public video sharing websites or sharing adult videos. Finally, a list of 59 websites are decided to be investigated as shown in Table 2. The following steps are performed to see whether media fragment sharing is available via user interface:

1. Open the landing page of a random video. If the website is not accessible for any reason, such as broken link and district restriction, we will mark the website as "Unknown".
2. On the landing page, find out whether there is any social sharing button allowing users to share the video at a certain time point.
3. Look for buttons or right click menu inside player indicating that a user can highlight a certain temporal or spatial area in the video.
4. Go to Twitter and search whether any video fragment has been shared.

If none of the above step leads to any clue about media fragment, we would make the conclusion that this video hosting website does not support media fragments, at least does not support W3C-MFURI. There are some obvious limitations in the methodology. Firstly, the methodology did not investigate,

[13] Accessed Oct, 2013, `http://en.wikipedia.org/wiki/List_of_video_hosting_services`

[14] `http://ted.com`

[15] `http://videolectures.net`

as a uploader of the video, whether one can annotate part of the video and save it as a URI, even though the URI may not be public. Secondly, some of the media fragment functions could be missed by the investigation, especially when there is language barrier and access restrictions. But as a preliminary study, experiment can still largely reflex the media fragment implementations for major video sharing platforms.

3.2 Survey Results

Table 2 shows the investigation results, where the **t** and **xywh** columns stand for the implementation of W3C-MFURI syntax for temporal and spatial dimensions. The value for each dimension is one of "Y", "P", "N", "U", which stand for fully implemented following the W3C-MFURI, partially implemented, not implemented and unknown respectively.

Table 2: Media Fragment Compatibility on Video Hosting Services (Oct. 2013)

Name	**t**	**xywh**	**Name**	**t**	**xywh**	**Name**	**t**	**xywh**
56.com	P	N	Archive.org	N	N	AfreecaTV	U	U
Blip.tv	N	N	BlogTV	N	N	Buzznet	N	N
Comedy.com	N	N	Crackle	N	N	DaCast	N	N
Dailymotion	P	N	EngageMedia	N	N	ExpoTV	N	N
Facebook	N	N	Funnyordie.com	N	N	Funshion	N	N
Fotki	N	N	GodTube	N	N	**Hulu**	P	N
Lafango	N	N	LeTV	N	N	Liveleak	N	N
Mail.ru	N	N	Mefeedia	N	N	Metacafe	N	N
Mevio	N	N	Mobento	N	N	Myspace	N	N
MyVideo	N	N	MUZU.TV	N	N	Nico Nico Douga	N	N
Openfilm	N	N	Photobucket	N	N	RuTube	N	N
Sapo Videos	N	N	SchoolTube	N	N	ScienceStage	N	N
Sevenload	N	N	SmugMug	N	N	Tape.tv	U	U
TED.com	N	N	Trilulilu	N	N	**Tudou**	P	N
Vbox7	P	N	Veoh	N	N	Vevo	N	N
Viddler	P	N	Videojug	N	N	Videolog	N	N
videolectures.net	N	N	Vidoosh	N	N	Viki.com	N	N
Vimeo	P	N	Vuze	N	N	Wildscreen.tv	N	N
Wistia	N	N	Yahoo! Video	N	N	**Youku**	P	N
YouTube	P	P						

As has been mentioned Section 2.1, YouTube allows users to share spacial fragment. However, the no spatial syntax is encoded in the URI and there is no affiliation between the fragment and the original video resource. So in Table 2, we indicate that YouTube has only partially implemented the spatial fragment.

In Table 2, 9 out of 59 websites partially implemented the notion of media fragment and there is only one implementation for spatial fragment. Table 3 lists some example URIs with media fragment encoded. Generally speaking, most of

them use URI query to encode the temporal fragment and only YouTube and Vimeo use URI hash. Only video in Hulu can encode both start and end time in the URIs. All of them adopt second as the basic unit to represent temporal scale, only YouTube also allows *ddhddmdds* string format.

Even though only a small portion of websites (9 out of 59) partially implemented the notion of media fragments, we still need to emphasise that the 9 websites may have covered most number of videos shared on the Web. We have not found any recent and valid resource that provides statistics about how many videos are shared on those websites, especially for large video sharing websites like YouTube, Vimeo and Dailymotion. But we can somehow have an empirical impression on this statistics based on the ranking or the unique monthly visitors of the those websites. According to an article published by eBizMBA[16] in April 2014, YouTube, Vimeo, Dailymotion and Hulu rank as 1st, 3rd, 5th and 6th among the top 15 of most popular websites. The sum of the estimated unique visitors per month from those four websites is nearly 73% of the total sum of the top 15 websites. So we can imply that the majority amount of videos we watch online could be shared on media fragment level.

Table 3. Media Fragment Syntax in Different Video Hosting Services (Oct. 2013)

Name	**Example url**
56.com	http://www.56.com/u92/v_OTgwMTk4NDk.html#st=737
Dailymotion	http://www.dailymotion.com/video/xjwusq&start=120 http://www.dailymotion.com/video/xjwusq?start=120
Hulu	http://www.hulu.com/embed.html? eid=sepr2dtbsyn7idlhbuzlbw&et=135&st=13
Vbox7	http://vbox7.com/play:cc7d3fc2?start=10
Viddler	http://www.viddler.com/v/bb2a72e9?offset=12.083
Vimeo	http://vimeo.com/812027#t=214 http://vimeo.com/812027?t=214
Tudou	http://www.tudou.com/listplay/H9hyQbAj4NM/2tzZHTtq4GA html?lvt=30
Youku	http://v.youku.com/v_show/id_XNjE2OTQ0MTI4.html? firsttime=147
YouTube	http://www.youtube.com/watch?v=Wm15rvkifPc#t=120 http://www.youtube.com/watch?v=Wm15rvkifPc?t=120 http://www.youtube.com/watch?v=Wm15rvkifPc&t=1h9m20s http://www.youtube.com/watch?v=Wm15rvkifPc#t=1h9m20s

From the survey results, we can see that only the Tweets containing URLs from those 9 websites are possibly valid. As Twitter is still banned in China (Mar, 2014), the indexer ignores those websites, and they are 56.com, Tudou and Youku. Hulu.com also has access restrictions from outside of U.S., so the

[16] `http://www.ebizmba.com/articles/video-websites`

indexer also ignores Hulu.com. Finally, YouTube, Dailymotion, Vbox7, Vimeo and Viddler are the video sharing applications that are selected to be monitored.

4 Indexing Media Fragments Using Twitter

This section will introduce the architecture of the Twitter Media Fragment Indexer and show some preliminary evaluation results. The main tasks of the indexer are: (1) collecting the Tweet messages filtered by the keywords corresponding to the video sharing platforms; (2) extracting URLs that encode media fragment information embedded in the message and (3) using the Media Fragment Indexing Framework to publish Tweet messages as annotations to the media fragment. We will also embed *VideoObject* defined in schema.org into the snapshot pages and investigate whether they can be recognised as videos in the Google search results. Even though the indexer has limited the scope of Tweets monitoring to only five websites, the number of Tweets is still considerably large. So as a proof-of-concept, the indexer only collects enough Tweets and media fragments to demo the system. The following subsections will detail the workflow of the indexer and discuss the evaluation results.

4.1 The Workflow of Twitter Media Fragment Indexer

Figure 3 shows the workflow of the Twitter Media Fragment Indexer. Generally, there are three stages: Twitter Media Fragment Crawling, Data Preparation and Media Fragment Indexing.

The first part (Process 1.1 in Figure 3) of the indexer is crawling the data from Twitter Stream API with "youtube, dailymotion, vimeo, vbox7, viddler" as the keywords to the "track" parameter in Twitter API 1.1. According to the Twitter status filter API[17], this filter phrase will return all the Tweets that match any of the keyword specified in the phrase regardless of the case. The matching works not only for the text attribute of the Tweet, but also the URLs in the expanded format, which by-passes the problem that shortened URLs usually do not match any of those keyword. When Tweets are returned by the Twitter Stream API, Process 1.1 will also filter out the Tweets that do not contain any URLs because the indexer is looking for URLs with media fragment information encoded. In Process 1.2, a programme is developed to parse the URLs contained in the Tweets based on the URL patterns observed in Table 3. The new parser is an extension of the Media Fragment URI Parser mentioned above. Basically, the new parser takes both the domain names and the query or hash string in the URLs into account, so that URLs from domains other than the five websites will not be accepted as media fragment URIs. The final process in the crawler is to save all the Tweets and the parsed media fragment information for the next stage.

After the required Tweets and media fragments have been saved locally, the Data Preparation stage will group the Tweets and media fragments by videos

[17] `https://dev.twitter.com/docs/api/1/post/statuses/filter`

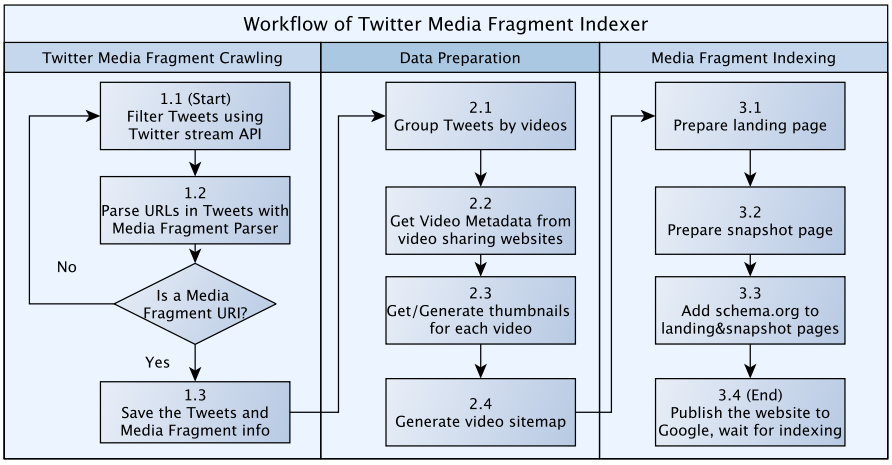

Fig. 3. The Workflow of Twitter Media Fragment Indexer

(Process 2.1) and collect necessary metadata for the content that will be displayed on landing pages and the snapshot pages. One video can be shared by many Tweets, and each Tweet may refer to a different time point in the video. So at the data preparation stage, the Tweets need to be grouped by video, so that Process 3.1 can assign a URI for each landing page, where different media fragments about the video can attach themselves at the back of the landing page URI using hashbang and the W3C-MFURI syntax.

In this demo, thousands of media fragments are expected to be created, so in order to let Google index the media fragments massively and automatically instead of submitting individual URLs for indexing, it is recommended to place a video sitemap at the root of the application. Following the guidelines of preparing a video sitemap for Google, some metadata about the video need to be retrieved from their original sharing platforms, which include the titles and descriptions (Process 2.2) and the thumbnails of the videos (Process 2.3). The thumbnails are downloaded and cached locally for search engine indexing. When all the information required for the video sitemap is ready, the sitemap.xml file will be generated containing every distinct media fragment URI crawled from Tweets. In some cases, especially in retweet messages, two or more Tweets may contain the URL about the same video at the same time point and all those Tweets will be annotating the same media fragment URI.

In stage 3, two sets of pages are prepared: landing page for users to view the video and Tweets, snapshot page for Google Ajax crawler to get the HTML content only related to a certain media fragment. Process 3.3 also embeds some simple Microdata into the landing and snapshot pages as the optimisation for the search engine. The final step is publishing the website and submit the sitemap to Google, and waiting for the snapshot pages to be crawled and indexed.

4.2 Implementation and Evaluation of Twitter Media Fragment Indexer

The demo website is available online[18] and any user can search the website content within Google using:

```
YOUR KEYWORDS site:twitter-mediafragment-monitor.herokuapp.com
```

and examine the indexed media fragments.

In the experiment, the crawling programme examined around 50 hours of non-stop Twitter stream (from 12:00:00 GMT, 22nd Dec, 2013 to 14:00:00 GMT, 24 Dec, 2013) with the filter phrase "youtube, dailymotion, vimeo, vbox7, viddler". During those 50 hours, the indexer examined 5,779,858 Tweets, in which 5,269,742 Tweets include one or more URLs. A media fragment URI parser has been developed for detecting the media fragments encoded in those URLs[19]. In total, there were 5,483,668 URLs processed by Process 1.2 in Figure 3, out of which 32,796 URLs are valid media fragment URIs and 32,754 Tweets contain valid media fragment URIs. So roughly, only 0.6% of the video URLs shared from those websites via Twitter encode media fragment information. Table 4 shows the breakdown number of the media fragment URIs shared in each website. YouTube takes nearly all the media fragment URIs shared on Twitter, while the indexer did not observe any media fragments shared from Vbox7 and Viddler. In the Data Preparation stage, the grouping of Tweets by videos (Process 2.1) result in 13,088 videos in total, which means at least one media fragment in those videos has been shared via Tweets. The number of videos are far fewer than the number of total Tweets with media fragment URIs. There are mainly two reasons. Firstly, many Tweets are sharing the same popular video, including the retweets. Secondly, Tweets are publicly available in all countries which have access to Twitter, however, some videos some of the video shared in Twitter are not accessible in UK. There are in total 104 videos fall into this access control. So the indexer will not be able to get the metadata of those videos and thus they are ignored in the final video collection.

Table 4. The Breakdown Number of Media Fragment URIs Shared in Each Website

Website	**Number of Media Fragment URIs**	**Percentage %**
YouTube	32,666	99.604
Dailymotion	101	0.308
Vbox7	0	0
Viddler	0	0
Vimeo	29	0.088

[18] `http://twitter-mediafragment-indexer.herokuapp.com`
[19] `https://github.com/yunjiali/Media-Fragments-URI-Loose`

In Process 2.3, 13,066 thumbnail pictures are retrieved from those websites and 22 thumbnail pictures are missing because the original pictures for some videos are not available. Finally, in Process 2.4, 17,854 video entries with media fragment URIs are generated in the sitemap. Even though there are 32,796 media fragment URIs collected from Process 1.3, some of them are referred to the same video and same time point. Sitemap should avoid duplicated URLs, even though they are shared by different Tweets. All the URLs included in the sitemap are newly mint within the "twitter-mediafragment-indexer.herokuapp.com" domain and the "content_loc" attribute in the video sitemap is used to link the landing page to original URL of the video.

For the Media Fragment Indexing stage, a very simple landing page and snapshot page are designed as a proof-of-concept (Figure 4). The temporal fragment is highlighted with the corresponding Tweet when the landing page is opened. *VideoObject* defined in schema.org are embedded in both landing page and snapshot page. At the time of writing this paper, 17,479 URLs in the sitemap out of

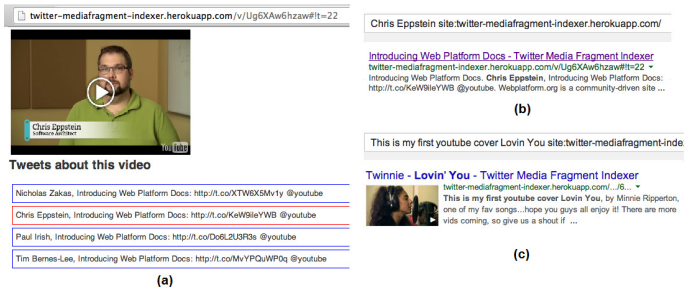

Fig. 4. The Landing Page in Twitter Media Fragment Indexer

the total 17,854 (around 97.9%) URLs have been indexed by Google as "Web pages" in the Google Webmaster Tools[20]. However, only 775 of the 17,854 URLs are indexed as "video" object. For the sake of evaluation, four Tweets are deliberately created during the experiment time as seen in Figure 4. All the four Tweets annotate the "Introducing Web Platform Docs" video on YouTube[21]: The URLs in the Tweets are media fragment URIs pointing to the times that the persons indicated in the Tweets start to talk in the video. For example, the second Tweet shows that Chris Eppstein is interviewed at the 22nd second in the video. As the evaluation results, those Tweets have been collected from the crawler and the media fragment URIs are included in the sitemap, which is submitted to Google. After those URIs are indexed by Google, searching one of the names in the Tweets in Google returns the "ugly URL" with hashbang. Clicking on that URL opens the landing page and the video player starts playing the

[20] `https://www.google.com/webmasters/tools/home?hl=en`
[21] `http://www.youtube.com/watch?v=Ug6XAw6hzaw`

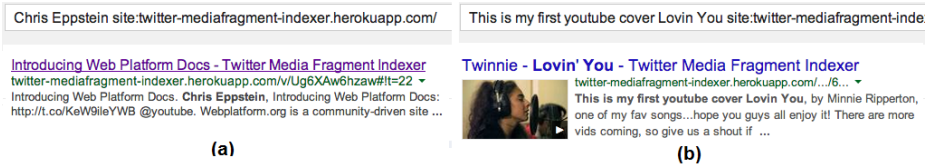

Fig. 5. Searching Media Fragment URIs Provided by Twitter Media Fragment Indexer

video from the start time represented by the media fragment URI in the corresponding Tweet. Figure 5(a) shows that searching "Chris Eppstein" will return the following URL:

`http://twitter-mediafragment-indexer.herokuapp.com/v/Ug6XAw6hzaw#!t=22`

Opening this link will lead to the landing page in Figure 4 and the video player will jump to the 22nd second of the video, while Tweet containing keywords "Chris Eppstein" is automatically highlighted. In theory, we can redirect further redirect the request to YouTube domain with the start time attached at the end of the YouTube URL, which is more user-friendly. But we did not implement this function because we want to show the Tweets related to the video in this demo. Figure 5(b) shows that some web page has been recognised as video, so a thumbnail is displayed with the search result. So the whole workflow suggested by Twitter Media Fragment Indexer can successfully use Twitter as a social annotation platform for media fragments indexing in Google.

5 Conclusion

To enable the media fragment indexing in larger scale, social media, such as Twitter, is proposed to be used to acquire more annotations linked to media fragments.The survey conducted in Section 3 has found out that most of the major video sharing platforms allow users to share a deep-linking to a certain time point of the video, so it is possible to monitor the sharing activities from social media and collect them as annotations as the media fragments. Based on the survey, Section 4 designs the Twitter Media Fragment Indexer to collect Tweets as annotations of media fragments and use the Media Fragment Indexing Framework to make the annotations shared on social media searchable in Google. The experiment result has shown that, after 50 hours' monitoring, YouTube is the most important resources of media fragment sharing on Twitter. 17,854 media fragment URIs have been created automatically from Twitter monitoring and more than 97% of them are successfully indexed by Google as web pages. 775 of them has been indexed as video objects thanks to the video sitemap and the embedded structured data.

While Twitter is used as the input data source of media fragment annotations, we still need further research on how valid the Tweets could be served as media fragment annotations. As can be seen in the evaluation, many Tweets do not

contain useful information for search and many retweets are simply repeating the same information. On the other hand, the methodology could be applied to other social media resources. For example, a programme can crawl subtitles or timed-text from YouTube and Dailymotion APIs and chunk the video accordingly for media fragment indexing. Nevertheless, the framework can also be easily applied for images and spatial dimensions.

Currently, there is no clear definition for media fragments as objects in the embedded semantic markups, such as schema.org. This leads to the difficulty of embedding rich RDF descriptions of media fragments. We can use the properties defined in schema.org for video object, but some special markups should be defined to address media fragments. In the future work, we are also planning to expand the period and the amount of Tweets that we crawl and further approve that this methodology can be applied for billions of Tweets. The social media content attached to media fragments can be processed using named entity recognition tools, which can be further used as the input parameter for to video classification [3].

References

1. Adida, B., Birbeck, M.: RDFa Primer (Oct 2008), `http://www.w3.org/TR/xhtml-rdfa-primer/`, http://www.w3.org/TR/xhtml-rdfa-primer/
2. Hickson, I.: HTML Microdata (Feb 2012), http://dev.w3.org/html5/md/
3. Li, Y., Rizzo, G., Redondo García, J.L., Troncy, R., Wald, M., Wills, G.: Enriching media fragments with named entities for video classification. In: Proceedings of the 22nd international conference on World Wide Web companion. pp. 469–476. International World Wide Web Conferences Steering Committee (2013)
4. Li, Y., Wald, M., Wills, G.: Let google index your media fragments. In: WWW2012 Developer Track (April 2012), `http://eprints.soton.ac.uk/336529/`
5. Troncy, R., Mannens, E., Pfeiffer, S., Deursen, D.V.: Protocol for media fragments 1.0 resolution in http (Dec 2011), http://www.w3.org/TR/media-frags/
6. Troncy, R., Mannens, E., Pfeiffer, S., Deursen, D.V.: Media fragments URI 1.0 (basic) (Mar 2012), http://www.w3.org/TR/media-frags/