A Tool for Monitoring of YouTube Content

Intzar Ali Lashari and Uffe Kock Wiil
The Maersk Mc-Kinney Moller Institute, University of Southern Denmark
Campusvej 55, DK-5230 Odense M, Denmark

Keywords: Social Media, YouTube, Monitoring, Information Visualization, Social Network Analysis and Mining.

Abstract: The expansion in use of social media has been very significant in the past decade. It has become a topic of interest for many researchers to find the most connected people, the most influential people, etc. in social media for various purposes. However, collection and monitoring of data in abundance from social media is difficult. This paper describes a new tool that can collect, monitor, and mine data from YouTube. The tool is part of a larger framework aimed at monitoring various social media including Facebook, Twitter, and YouTube. A specific case focusing on "Islamic Jihad Holy War" demonstrates the features of the tool.

1 INTRODUCTION

A rapidly increasing number of people share information with others by using different social media sites such as Facebook, Twitter, YouTube, LinkedIn, MySpace, etc. A social media site provides a platform to the users for the social relations based on their respective interests, relations, and activities which they want to share with their social circle on the internet. Social networks can be described as information networks in which actors are represented by nodes and relations are represented by edges (Aggarwal, 2011). The context of a social network works as a motivation for the actors in the social network and the resulting content generated by the actors and the structure encourages the participation and subsequently affects the social relation (Zeng and Wei, 2013). The increase in use of social media has created an interest in the research community to analyse and mine the social media data. However, social network analysis and mining is becoming more and more challenging as the generated content becomes richer and more abundant (Aggarwal, 2011).

Social media is used for many different purposes. Three examples are given here: (1) In countries like Egypt, Tunisia, and Yemen, rising action plans such as protests made up of thousands, have been organized through social media such as Facebook, YouTube, and Twitter. They used Facebook to schedule the protests, Twitter to coordinate, and YouTube to tell the world as part of the Arab Spring uprisings (2012) (Polymic, 2012). (2) Social media was extensively used during the East Japan Earthquake (2011) to share information about the disaster and getting in touch with missing relatives (Telegraph, 2011). (3) Social media is used increasingly for militant islamist propaganda with the intent to radicalise Muslims (CTA, 2012). Hence, their is an increasing interest in the ability to monitor the social media for instance to get information about evolving events and in the interest of public safety.

In this paper, we present a tool for monitoring of YouTube content. The tool is part of a larger framework aimed at monitoring social media such as Facebook, Twitter, and YouTube. On YouTube various types of videos can be uploaded by different users. Users can watch, like/dislike, and comment upon the available videos. It is possible to see how many times a video has been viewed as well as other metrics about videos and comments. Data changes over time as people engage actively on YouTube. Section 2 describes YouTube in more detail. A set of monitoring metrics are proposed that can help find the most influencial people, videos, etc. on YouTube based on social network analysis and mining. Section 3 describes related work. Section 4 presents the tool and Section 5 focuses on a specific case that is used to demonstrate the features of the tool. Finally, Section 6 concludes the paper.

2 YOUTUBE

What is the nature of user activity on YouTube? It is

Ali Lashari I. and Kock Wiil U..
A Tool for Monitoring of YouTube Content.
DOI: 10.5220/0005068401710178
In *Proceedings of the International Conference on Knowledge Management and Information Sharing* (KMIS-2014), pages 171-178
ISBN: 978-989-758-050-5

an assumption by some scholars (Tannen, 1999) that in on-line environments the prevalence of anonymity directly spawns antagonism, and that an increase in identity information will decrease the communicative hostility. However some scholars hold the opposing view that additional identity information, such as, facial and bodily information, does not guarantee cordial communication among the commentators (Lange, 2007).

On-line communicative environments like video sharing sites, blogs, etc. provide a platform for exchange of comments. The comments can have a wide range of reaction, ranging from ecstatic praise to extreme hate to threat of physical violence. However, given the fact that on-line interactions take place in relative anonymity, not all participants may view certain critical comments as a problem that requires greater regulatory control, which can be viewed as a threat to limit participation (Lange, 2007).

Two major types of negative comments can be distinguished—comments that are hateful or threaten violence and comments that provide constructive criticism. Hateful comments are often the cause of discouragement of open self-expression on the site. The driving component in such on-line exchange of comments are the users of video exchange site, blogs, etc. who participate in the textual communication. According to a previous study, the participants can be classified in one of several categories (Lange, 2007):

1. **Former Participants.** They are those who no longer post videos, blog posts, etc. but still maintain an account.
2. **Casual Users.** They typically don't have an associated account. However, they tend to view videos or read blogs, etc. when they wish to search for something while surfing the Web, and are prompted to view a video or read a blog post.
3. **Active Participants.** They usually have an associated account, and occasionally participate by uploading videos, writing posts, and/or by commenting on other people's contributions.
4. **Highly-active Participants.** These have a more intense level of activity and participation on the social sites, spend much more time regularly uploading content and maintaining their sites. They tend to promote their work within and outside of the content platform.
5. **Celebrities.** Similar in many respects to the last category, but are also well-known despite their online presence in the form of YouTube channels, blogs, FaceBook pages, or Twitter handlers. They are often in a position to influence discourse by the content they upload/create, and other interactions on such sites.

The above categories only provide a description of the relative levels of participation among the whole body of users and are not mutually exclusive.

In order to monitor user activity on YouTube, it is necessary to think in terms of metrics that can quantify users and their activities. Hence, important questions are: What would we like to monitor? And what can be monitored given the available YouTube API (YouTube, 2014)? In terms of videos, it is interesting to see how influential they are. This can be measured by the number of views, the number of likes/dislikes, and the number of comments. In terms of users, it is interesting to see how influential they are and who they engage with. The former can be determined by the number of videos they post and how influential they are (see above). The latter can be found by analyzing the social network formed by the activities of users based on who comments on what videos.

For the purpose of this work, we focus on videos that are retrieved based on keyword-matching using YouTube's API. For each matching video, data about users who have made comments on that video are collected. The user data collected includes the YouTube identifiers for users who uploaded the video and users who subsequently commented on the video. This enables us to generate a social network of YouTube users from a contextual perspective, where the context is defined by keywords typed in by the person (investigator or analyst) that wishes to monitor a certain event, activity, etc.

3 RELATED WORK

Internet-based applications that are build on the technological and ideological foundation of Web 2.0 allows users to create and exchange the contents of their interest (Kaplan and Haenlein, 2010). These applications include Facebook, Twitter, and YouTube to name a few.

It has been suggested that videos are a very potent medium for affecting the attitudes and political will of the intended audience (Farwell, 2010). They can be used to communicate a message for influencing values, culture, attitudes, and opinions. Online video platforms, like YouTube, provide a very effective and cheap way to reach mass audiences that would otherwise be difficult to reach using conventional means.

Terrorist networks has received much attention after 9/11 (2001). The Al-Qaeda network and related Jihadist organizations have been analyzed with respect to their Internet based information strategies. Related

work in this area has considered online content of Jihadists and their supporters (Conway, 2006) with a particular emphasis on the content of Jihadi videos including various types of videos and their impact on their audiences (Salem et al., 2006), (Kimmage and Ridolfo, 2007), (Salem et al., 2008).

For example, the HBO documentary called Baghdad ER (Baghdad, 2006), which dealt with the subject of providing emergency medical care to wounded US personnel in the battlefield, was re-created by Al-Qaeda based in Mesopotamia by replacing the original soundtrack with their own, and by making an entirely different beginning and ending to show that the US forces are sustaining loses and being defeated in combat.

Similarly, in Iraq and Afghanistan, terror groups use videos to demonstrate their victories over the opposing side. Viewer-enthusiasm is gauged by how quickly the video spreads over the internet and into news media sites (Farwell, 2010). Because of this unconventional approach, they have a greater chance of getting news coverage via satellite TV which has a large number of viewers in the Arab world and similar conflict zones.

The main idea is to identify the impact of videos to the audiences. Funders and policy makers have shown an increased interest in learning the ways of violent radicalization (Council of the European Union, 2005). A high profile example is Hussain Osman, one of the London bombers, who claimed to have been influenced by watching videos of the conflict in Iraq along with reading about jihad in an online forum.

In this paper, we create a network of YouTube video uploaders to investigate the possibilities of radicalization via the Internet specifically from YouTube as opposed to analysis of jihadist sites. We aim to find the users that are registered on YouTube and have uploaded videos and/or have commented on video content related to a given context. Hence, our work is closely related to the work of (Chen et al., 2003; Wen et al., 2007; Das et al., 2008).

With respect to data collection, analysis, and visualization, Coplink (Chen et al., 2003) was one of the first systems to successfully address the domain of criminal network investigation. The system was first deployed at the Tucson Police Department. The system collects, combines, and analyzes data from various sources and generates overviews of the information for the investgators to help them solve cases.

With respect to the idea of context, (Wen et al., 2007) presents an intelligent information system that performs an investigation task for detecting frauds. The authors have contributed by developing two notions: 1. Context and 2. Context-awareness. Related to context, the paper defines the term investigation context and with regard to context-awareness the investigator can adaptively retrieve data and evaluate the relevant information for the ongoing investigation.

With respect to collection of user networks from social media, (Das et al., 2008) focuses on improving performance in information collection of a social graph of users' neighbours in a dynamic social network. In the study, the author has introduced a sampling based algorithm for quickly approximating quantities of interest, the vicinity of a user's social graph that explores the variants of correlation across the sample. The algorithm can be used to rank the items in the neighbourhood of a user.

Pippal et al. (Pippal et al., 2014) provide a recent survey of data mining approaches and methods in social networking sites, including micro-blogging, twitter, YouTube, instagram, blogs, forums, etc. Closely related work is done by Agarwal and Sureka (Aggarwal and Sureka, 2014) where the authors analyse YouTube metadata for privacy invading and harrassment content. He et al. model user comments on YouTube videos as a bipartite graph to predict the popularity of videos and other item on the Web. (He et al., 2014)

The purpose of our study is to develop a tool for context-based monitoring of social media data based on a set of monitoring metrics. To our knowledge, no one has defined monitoring metrics for YouTube with the intent to develop a tool to explore and monitor influential videos, users, and networks.

4 THE YOUTUBE MONITORING TOOL

The YouTube monitoring tool is a part of our framework entitled "Keyword-based Social Network Analysis Framework" (KSNAF). KSNAF aims at supporting collection, monitoring, and mining of social media data from a contextual perspective. Overall, the framework must meet the following requirements:

1. The framework enables context-based search. In the case of YouTube, the framework provides facilities to collect videos that match a given context based on keywords defined by the investigator.
2. The framework determines relationships among the users that are retrieved as the result of a context-based search. In the case of YouTube, the framework can build a social network based on the users' activities (comments on videos).
3. The framework enables investigators to view data related to a given set of monitoring met-

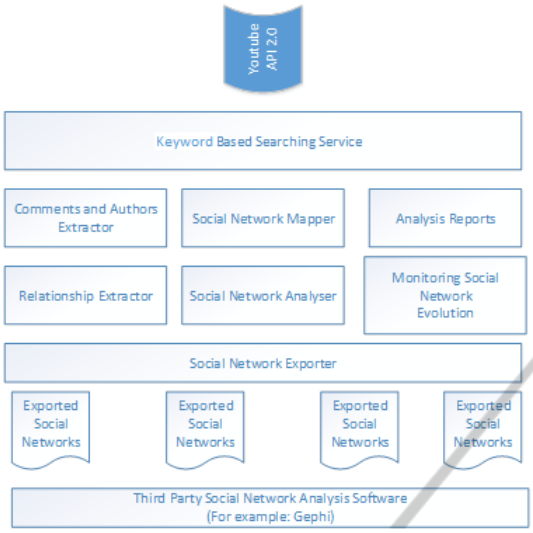

Figure 1: System architecture of KSNAF.

rics. In the case of YouTube, the framework can point to the most influential people, videos, and (sub)networks.

4. The framework supports social network analysis related to the generated social network.
5. The framework can export data in formats that are importable by various social network analysis, mining, and visualization packages.

An architectural overview of the KSNAF framework is shown in Figure 1. KSNAF includes the following components to meet the above described requirements.

- **Comments and Authors Extractor.** This component is responsible for extracting commenters and authors with respect to a particular video under investigation.
- **Relationship Extractor.** This component is responsible for extracting relationships as ordered triplets of the form (Author, Commenter, Comment).
- **Social Network Mapper.** This component is responsible for mapping the relationships extracted by the Relationship Extractor component into a social network. This component contains the coordinating logic that is used in interaction among other components. At present, the component uses Algorithm 1 to generate a context-aware social network.
- **Social Network Analyser.** This component carries out social network analysis on the networks generated by the Social Network Mapper component. It supports a number of network analysis algorithms, computing various metrics for the level of nodes, links, and the whole network. It comes up with results such as what is the most frequent Author-Commenter pattern with respect to the searched context. Examples of network metrics computed by this component include mean path length, density, and centrality measures (like degree, betweenness, closeness, etc.). Finally, key players, weighted link analysis (Memon, 2012) and determining various clusters in the resulting social network is supported by this component.
- **Analysis Reports.** This component generates different reports on the basis of the Social Network Analyser components' findings over a particular span of time.
- **Monitoring Social Network Evolution.** This component is responsible for determining the changes over time in the structure obtained from the user-generated activity of writing comments and/or uploading videos. Hence any changes in the network are identified and the network is updated accordingly. These updates are done over time by the Social Network Analyser component in a particular search context.
- **Social Network Exporter.** This component exports the social network in a variety of formats to enable social network analysis, mining, and visualization of third-party software packages. At present, the component supports comma separated values (CSV), XML, and GraphML formats.

In Algorithm 1, V denotes the collection of all videos, D gives the maximum allowable depth for search, and d denotes the current search depth, A denotes the collection of authors who have commented on the relevant videos, and finally G denotes the network graph.

5 CASE STUDY AND RESULTS

In this section, we present a case study to demonstrate the features of the developed YouTube monitoring tool. The selected case is related to the uprising (civil war) in Syria. According to the Danish Security and Intelligence Service as well as Danish media (TV2, 2013), (DR, 2014), (Politiken, 2013), around 90 Muslims have left Denmark to join the "holy war" in Syria. In particular, one video on YouTube is heavily criticized for encouraging Danish Muslims to join the civil war in Syria. The video is entitled "A Danish terrorist in Syria!" (http://www.youtube.com/watch?v=VTJ1ynW60gU). This clearly demonstrates the need for a monitoring tool like the one described in this paper.

Data:
Videos V retrieved as result of searching with context C on YouTube
D is the maximum allowed depth at for searching
d is the current depth
Result:
Graph G of relationships on YouTube
for *each video vx in V* **do**
  Retrieve the comment authors A;
  Retrieve the comments C;
  for *each commenter cx in C* **do**
    Add an edge that connects A with cx in G;
    Retrieve the videos vx posted by Cx;
    if $d < D$ **then**
      Add 1 to d;
      Execute the Algorithm 1 with arguments Vx, D, d, G;
    else
      Return G
    end
  end
end

Algorithm 1: Searching the YouTube graph.

We have used the case study to validate the framework. Basically, the tool is generic and any keyword-based search of related videos and other associated data, such as, user IDs, user locations, comments, etc. can be retrieved from YouTube. The data which we have collected from YouTube using the available API (YouTube, 2014) is based on the context "Islamic Jihad Holy War.

We provide an analysis of the corresponding social network that was generated based on the data collected from YouTube. Figure 2 depicts the results of searching the context and yields the videos that are related to the context. The resulting videos are then further analysed to determine what are the commenters of the retrieved videos and what are other videos posed by the commenter. Such analysis is beneficial to determine the regular commenter on the videos in the given context, to determine the key players and sources that positively contributes in uploading the most viewed or commented videos, and finally to reveal interesting communication patterns like upload-comment-upload. Figure 3 shows the different social networks behind each of the videos that is retrieved against the context search. Table 1 shows the location of the commenters.

We have run the tool using a Windows 7 PC having the following configuration: Intel (R) Core (TM) 2 CPU6600 2.40GHZ and 4.00GB RAM. It took 6

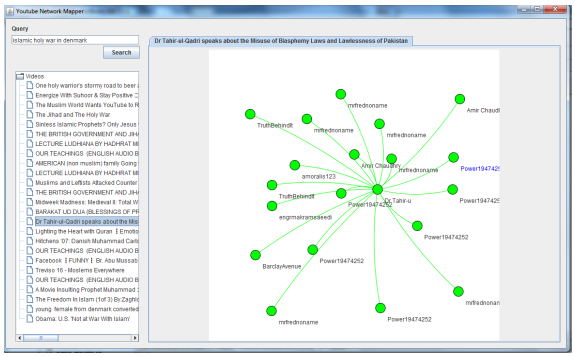

Figure 2: Context-based search of "Islamic Jihad Holy War".

Table 1: Location of YouTube commenters.

Country of Users	No. of Users
United States (US)	152
Great Britain (GB)	13
Germany (DE)	2
Sweden (SE)	6
Canada (CA)	16
Australia (AU)	6

minutes and 12 seconds to collect the data of authorship and generate the commenter clustering. It will take more time for the collection of data if the depth of the users' interconnection is increased. For this case we have used a depth of 5.

We have used the visualization tool Gephi and the Yifan Hu graph drawing algorithm (Hu, 2005) which is efficient and high quality in visualization of relationships. Figure 4 shows the users as nodes and the relationships among them, as links, connecting users who have commented on videos by other users. Although, the networks show different snapshots collected on different dates, the nodes in all the networks have a one-to-one correspondence among themselves.

Figure 4b-f shows the evolution over time in the network from March 20 to April 13. Figure 4a shows the network on April 15. Hence, Figure 4 demon-

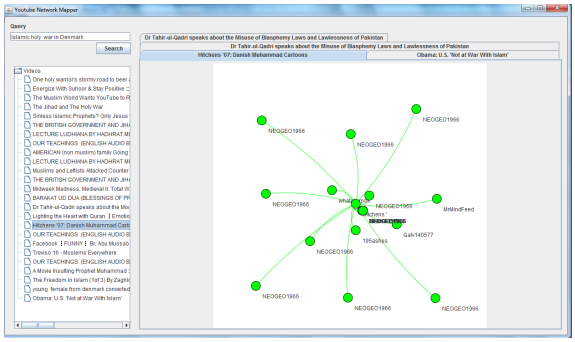

Figure 3: Multiple social networks extracted in the context of the query.

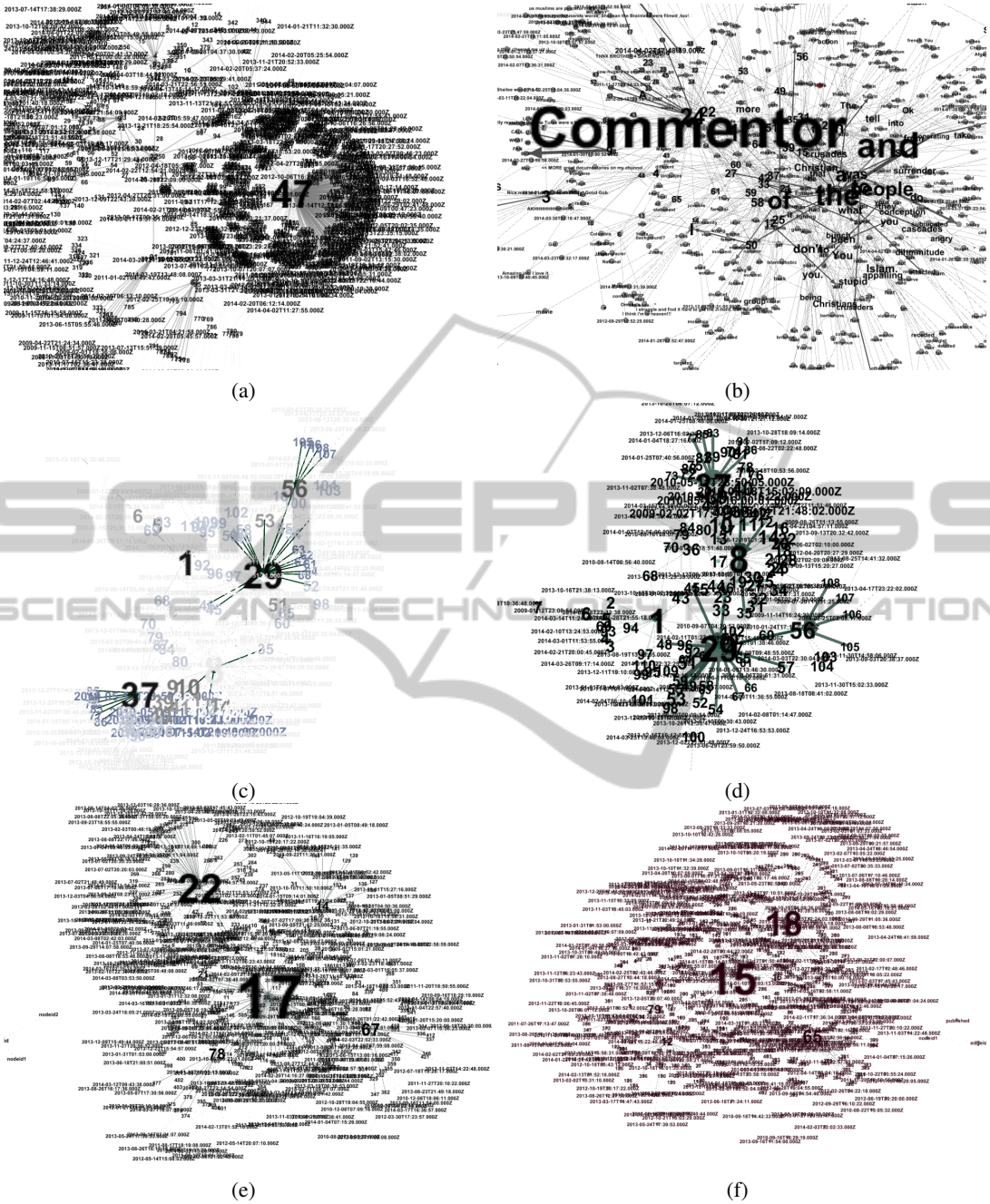

Figure 4: Network graphs for the extracted YouTube data. (a) shows the entire network (April 15, 2014), (b)–(f) networks collected from March 20 to April 13, 2014 (newest first).

strates that we are able to monitor (track) how the influence of users and videos change over time. The highly central users are those whose videos have generated most comments by other users (e.g., users 47, 92, 49, 86, 22, 91, etc.) as shown in Table 2. It is interesting to observe that the highly central users are connected among themselves, i.e., not only does their own videos generate a large volume of user-comments (they are popular and influential) but such users are commenting on other popular videos as well. Moreover, Table 3 shows the top most and least popular users in terms of the number of video views they have received from other users in the network.

For the collection of data from YouTube we have used three different API's provided by YouTube: (1) data collection based on keyword search, (2) data col-

Table 2: Comments on videos by highly central users.

User ID	YouTube name	No. of comments
47	Jman92854	734
92	meteosurreal	83
49	xHolyCrusader	64
86	Indago55	58
22	Phoebe Igor	55
91	wadeywilson101	48

Table 3: Users with highest and lowest view counts.

User ID	Highest view count	User ID	Lowest view count
111	158748	95	3
106	77002	175	12
116	50239	45	33
59	19310	174	35
120	19186	19	38
25	16631	27	38
88	7876	121	43
135	5501	169	44
177	5468	99	44
79	4861	12	46

lection on content authors, and (3) data collection on author profiles. For the information collection of data of the network we have connected the users information on node ID which is the user's YouTube ID and related commenters data depending on the depth of the relations on comments. There is not much variation in the location of the users from dataset. The majority of commenters belong to North America (US and Canada). Figure 4b, the commenter with node ID 1 has claimed that "The Muslims dont need Crusaders to embrace jihad, jihad is fundamental to Islam" This is one of the examples of comments obtained from the dataset based on the mentioned keywords. The developed tool has the main purpose of supporting user-defined keyword-based collection of data from YouTube for monitoring.

6 CONCLUSION

The main goal of this study is to find YouTube users who is uploading videos related to the chosen context and those who have commented on those videos. Our tool is able to extract data related to YouTube videos based on keywords chosen by the investigator. The collected data is stored in a relational database for later operations and queries. The collected data includes user attributes, user comments, location, timestamps, etc. As far as our tool is concerned, we are able to find relations of YouTube video content and present them as a social network graph that shows the relationships. In addition, various statistical metrics of the social network graph are presented. These metrics show the connectedness and influence of users. Our tool is able to detect and update the relationships on a regular basis and show that on a specific time who is closely connected to whom and who is interested in whose content and what content.

As part of the near term future work, additional case studies will be made to further validate the YouTube monitoring tool. Also, we wish to further investigate how to best visualize the evolution of the collected metrics and networks over time.

Monitoring tools for Facebook and Twitter are being developed in parallel to the described tool. They will be documented in subsequent publications.

REFERENCES

Aggarwal, C. C. (2011). *An introduction to social network data analytics*. Springer.

Aggarwal, N. and Sureka, A. (2014). Mining youtube metadata for detecting privacy invading harassment and misdemeanour videos. Master's thesis, Indraprastha Institute of Information Technology, Dehli, India.

Baghdad (2006). Baghdad er. http://www.imdb.com/title/tt0802944/.

Chen, H., Schroeder, J., Hauck, R. V., Ridgeway, L., Atabakhsh, H., Gupta, H., Boarman, C., Rasmussen, K., and Clements, A. W. (2003). Coplink connect: information and knowledge management for law enforcement. *Decision Support Systems*, 34(3):271–285.

Conway, M. (2006). Terrorism and the internet: New media new threat? *Parliamentary Affairs*, 59(2):283–298.

Council of the European Union (2005). The european union counter-terrorism strategy. Technical report, Council of the European Union Report.

CTA (2012). Center for terror analysis danish security and intelligence service. militant islamist propaganda on facebook and youtube (in danish). https://www.pet.dk/CTA/~/media/CTA/UKLCTAanalyseFacebookogYoutubepdf.ashx.

Das, G., Koudas, N., Papagelis, M., and Puttaswamy, S. (2008). Efficient sampling of information in social networks. In *Proceedings of the 2008 ACM workshop on Search in social media*, pages 67–74. ACM.

DR (2014). Danish television. 22 young muslisms from aarhus has left for syria (in danish). http://www.dr.dk/Nyheder/Ligetil/Dagens_fokus/Indland/2014/01/22_unge_muslimer_fra_Aarhus_er_rejst_til_Syrien.htm.

Farwell, J. P. (2010). Jihadi video in the war of ideas. *Survival*, 52(6):127–150.

He, X., Gao, M., Kan, M.-Y., Liu, Y., and Sugiyama, K. (2014). Predicting the popularity of web 2.0 items based on user comments. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 233–242, New York, NY, USA. ACM.

Hu, Y. (2005). Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, 10(1):37–71.

Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite! the challenges and opportunities of social media. *Business horizons*, 53(1):59–68.

Kimmage, D. and Ridolfo, K. (2007). Iraqi insurgent media: The war of images and ideas. *Radio Free Europe/Radio Liberty Special Report*, page 4.

Lange, P. G. (2007). Commenting on comments: Investigating responses to antagonism on youtube. In *Annual Conference of the Society for Applied Anthropology. Retrieved August*, volume 29, page 2007. Citeseer.

Memon, B. R. (2012). Identifying important nodes in weighted covert networks using generalized centrality measures. In *2012 European Intelligence and Security Informatics Conference*, pages 131–140. IEEE.

Pippal, S., Batra, L., Krishna, A., Gupta, H., and Arora, K. (2014). Data mining in social networking sites: A social media mining approach to generate effective business strategies. *International Journal of Innovations & Advancement in Computer Science*, 3(2).

Politiken (2013). Danish newspaper. danish muslisms are encouraged to holy war in syria (in danish). http://politiken.dk/indland/ECE2047926/danske-muslimer-opfordres-til-hellig-krig-i-syrien/.

Polymic (2012). Twitter revolution: How the arab spring was helped by social media. http://www.policymic.com/articles/10642/twitter-revolution-how-the-arab-spring-was-helped-by-social-media.

Salem, A., Reid, E., and Chen, H. (2006). Content analysis of jihadi extremist groups videos. In *Intelligence and Security Informatics*, pages 615–620. Springer.

Salem, A., Reid, E., and Chen, H. (2008). Multimedia content coding and analysis: Unraveling the content of jihadi extremist groups' videos. *Studies in Conflict & Terrorism*, 31(7):605–626.

Tannen, D. (1999). *The argument culture: Stopping America's war of words*. Ballantine Books New York.

Telegraph (2011). The telegraph. japan earthquake: how twitter and facebook helped. http://www.telegraph.co.uk/technology/twitter/8379101/Japan-earthquake-how-Twitter-and-Facebook-helped.html.

TV2 (2013). Tv2 news (denmark). terrorism expert: Danish jihad video is dangerous (in danish). http://nyhederne.tv2.dk/article.php/id-70767751:terroreksperT-dansk-jihadvideo-er-farlig.html.

Wen, Z., Zhou, M. X., and Aggarwal, V. (2007). Context-aware, adaptive information retrieval for investigative tasks. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 122–131. ACM.

YouTube (2014). Youtube api v3. https://developers.google.com/youtube/v3/.

Zeng, X. and Wei, L. (2013). Social ties and user content generation: Evidence from flickr. *Information Systems Research*, 24(1):71–87.