A Focused Crawler for Mining Hate and Extremism Promoting Videos on YouTube

Swati Agarwal, Ashish Sureka
Indraprastha Institute of Information Technology, Delhi (IIIT-D)
New Delhi, India
swatia@iiitd.ac.in, ashish@iiitd.ac.in

ABSTRACT

Online video sharing platforms such as YouTube contains several videos and users promoting hate and extremism. Due to low barrier to publication and anonymity, YouTube is misused as a platform by some users and communities to post negative videos disseminating hatred against a particular religion, country or person. We formulate the problem of identification of such malicious videos as a search problem and present a focused-crawler based approach consisting of various components performing several tasks: search strategy or algorithm, node similarity computation metric, learning from exemplary profiles serving as training data, stopping criterion, node classifier and queue manager. We implement a best-first search algorithm and conduct experiments to measure the accuracy of the proposed approach. Experimental results demonstrate that the proposed approach is effective.

Categories and Subject Descriptors

H.5.4 [**Hypertext/Hypermedia**]: Navigation; K.4.2 [**Social Issues**]: Abuse and crime involving computers; D.2.8 [**Metrics**]: Performance measures; H.3.1 [**Content Analysis and Indexing**]: Linguistic processing

Keywords

Social Media Analytics; Focused Crawler; Hate and Extremism Detection; Video Sharing Website; Online Radicalization.

1. RESEARCH MOTIVATION AND AIM

Research shows that YouTube has become a convenient platform for many hate and extremist groups to share information and promote their ideologies. The reason is because video is the most usable medium to share views with others [1]. Previous studies show that extremist groups put forth hateful speech, offensive comments and messages focusing their mission [3]. Social networking allows these users (uploading extremist videos, posting violent comments, subscribers of these channels) to facilitate recruitment, gradually reaching world wide viewers, connecting to other hate promoting groups, spreading extremist content and forming their communities sharing a common agenda [2] [6]. The presence of such extremist content in large amount is a major concern for YouTube moderators (to uphold the reputation of the website), government and law enforcement agencies (identifying extremist content and user communities to stop such promotion in country). However, despite several community guidelines and administrative efforts made by YouTube, it has become a repository of large amounts of malicious and offensive videos [5]. Detecting such hate promoting videos and users is a significant and technically challenging problem. 100 hours of videos are uploaded every minute, that makes YouTube a very dynamic website. Hence, locating such users by keyword based search is overwhelmingly impractical. The work presented in this paper is motivated by the need of a solution to combat and counter online radicalization. We frame our problem as: identifying such videos promoting hate and extremism on YouTube. The research aim of the work presented in this paper is to investigate the application of a focused crawler (best-first search) based approach for retrieving YouTube user-profiles promoting hate and extremism. To investigate the effectiveness of contextual features such as the title of the videos uploaded, commented, shared, and favourited for computing the similarity between nodes in the focused crawler traversal and to examine the effectiveness of subscribers, featured channels and public contacts as links between nodes.

2. BEST-FIRST SEARCH CRAWLER

The proposed method is a multi-step process primarily consists of three phases, Training Profile Collection, Statistical Model Building and Focused Crawler. We perform a manual analysis and a visual inspection of activity feeds and contextual metadata of various YouTube channels. We collect 35 positive class channels (promoting hate and extremism) used as training profiles. We build our training dataset by extracting the discriminatory features (user activity feeds such as titles of videos uploaded, shared, favourited & commented by the user and profile information) of these 35 channels using YouTube API[1]. We build a statistical model from these training profiles by applying character n-gram based language modeling approach. We build a focused crawler (best-first search) which is a recursive process.

[1]https://developers.google.com/youtube/getting_started

HT'14, September 1–4, 2014, Santiago, Chile.
ACM 978-1-4503-2954-5/14/09.
http://dx.doi.org/10.1145/2631775.2631776 .

Algorithm 1: Focused Crawler- Best First Search

```
   Data: Seed User U, Width of Graph w, Size of Graph s, Threshold th, N-gram Ng, Positive
         Class Channels U_p
   Result: A connected directed cyclic graph, Nodes=User u
1  for all u ∈ U_p do
2  |  D.add(ExtractFeatures(u))
   end
   Algorithm BFS(U)
3  |  while graphsize < s do
4  |  |  userfeeds U_f ←ExtractFeatures(U)
5  |  |  score score ←LanguageModeling(D, U_f, Ng)
6  |  |  if (score <th) then
7  |  |  |  U.class ←Irrelevant
   |  |  else
8  |  |  |  U.class ←Relevant
9  |  |  |  Hashmap U_sorted·InsertionSort(U, score)
   |  |  end
   |  |  for i ← 1 to w do
10 |  |  |  Hashmap U_graph·add(U_sorted(i))
   |  |  end
11 |  |  for all U_g ∈ U_graph do
12 |  |  |  fr = Extract_Frontiers(U_g)
13 |  |  |  Hashmap U_crawler·add(fr)
   |  |  end
14 |  |  for all U_fr ∈ U_crawler do
15 |  |  |  BFS(U_fr)
   |  |  end
   |  end
```

Algorithm 2: Frontier Extraction for a YouTube User

```
   Data: User u
   Result: Frontiers of a channel
   Algorithm Extract_Frontiers(U)
1  |  u_subs ←u.getSubscribers()
2  |  u_fc ←u.getFeaturedChannels()
3  |  u_con ←u.getFriends()
```

It takes one YouTube channel as a seed (a positive class channel) and extract it's contextual metadata (user activity feeds and profile information) using YouTube API. We find the extent of textual similarity between these metadata and training data by using statistical model (build in phase 2) and LingPipe API [2]. We implement a binary classifier to classify a user channel as relevant or irrelevant. A user channel is said to be relevant (hate and extremism promoting channel) if the computation score is above a predefined threshold. If a channel is relevant, then we further extend it's frontiers (links to other YouTube channels) i.e. the subscribers of the channel, featured channels suggested by the user and it's contacts available publicly. We extract these frontiers by parsing users' YouTube homepage using jsoup HTML parser library[3]. We execute focused crawler phase for each frontier recursively which results a connected graph, where nodes represent the user channels and edges represent the links between two users.

Inputs to the algorithm is a seed (a positive class user) U, width of graph w i.e. maximum number of children of a node, size of graph s i.e. maximum number of nodes in graph, threshold th for classification, n-gram value Ng for similarity computation (language modeling), and a lexicon of 35 positive class channels U_p. We compare each training profile with all profiles and compute their similarity score for each mode. We take an average of these 35 scores and compute the threshold values. The proposed method (Algorithm 1) follows the standard best-first traversing to explore relevant user to seed input. Best-First Search examines a node in the graph and finds the most promising node among it's children to be traversed next [4]. This priority of nodes (users) is decided based upon the extent of similarity with the training profiles. A user with the similarity score above a specified threshold is said to be relevant and allowed to be extended further. If a node is relevant and has the highest priority (similarity score) among all relevant nodes then we extend it first and explore it's links and discard irrelevant nodes. We process each node only once and if a node appears again then we only include the connecting edge in the graph. Steps 1 and 2 extract all contextual features for 35 training profiles using a feature extraction algorithm and build a training data set. Steps 4 and 5 extract all features for seed user U and compute it's similarity score with training profiles using character n-gram and language modeling. Steps 6 to 8 represent the classification procedure and labeling of users as relevant or irrelevant depending upon the threshold measures.

BFS method has non-binary priority values assigned to each node. The priority values are the similarity score, which is computed by comparing the users' contextual metadata (user activity feeds and profile information) with training profiles. Steps 9 and 10 make a list of top w (maximum number of children, a node can have) users among relevant users based upon their similarity score, sorted in a decreasing order. Steps $11-13$ extracts frontiers of a user channel using Algorithm 2. Steps 14 and 15 repeat Steps $3-13$ for each frontier extracted. We execute this function till we get a graph with desired number of nodes or there is no more node is left to extend.

Table 1: Best-First Search Confusion Matrix

		Predicted	
		Relevant	**Irrelevant**
Actual	**Relevant**	921	314
	Irrelevant	125	67

3. PERFORMANCE EVALUATION

The crawler requires exemplary documents or training examples to learn the specific characteristics and properties of documents in the training dataset. A statistical model (text classifier) needs to be built from a collection of documents pertaining to a predefined topic. We create a list of 35 user-ids used as training profiles. The 35 user ids consists of 612 videos and hence the training is performed on 612 videos. We obtain the training dataset by manually searching (keyword based) for anti-India hate and extremism promoting channels using YouTube search and traversing related video links (using the heuristic that videos on similar topic will be connected as relevant on YouTube). We select 10 random positive class (hate and extremist) channels for creating test dataset. Each user works as a seed input to the focused crawler. To evaluate the effectiveness of our solution approach we execute our focused crawler several times for various configurations and seed. Table 1 shows the confusion matrix for binary classification performed during Best-First Search approach. Given the input of 10 seed users and 6 modes (pair of threshold and n-gram values) we get different number of connected users in each iteration. To measure the accuracy of our proposed approach we collect results of all 60 iterations and classify 1046 (921 + 125) users as relevant and 381 (314 + 67) as irrelevant users. There is a misclassification of 25.42% and 65.10% in predicting the relevant and irrelevant users respectively.

[2]http://alias-i.com/lingpipe/index.html
[3]http://jsoup.org/apidocs/

4. REFERENCES

[1] Hsinchun Chen, Dorothy Denning, Nancy Roberts, Catherine A. Larson, Ximing Yu, and Chunneng Huang. The dark web forum portal: From multi-lingual to video. In *ISI*, pages 7–14. IEEE, 2011.

[2] Maura Conway and Lisa McInerney. Jihadi video and auto-radicalisation: Evidence from an exploratory youtube study. In Daniel Ortiz-Arroyo, HenrikLegind Larsen, DanielDajun Zeng, David Hicks, and Gerhard Wagner, editors, *Intelligence and Security Informatics*, volume 5376 of *Lecture Notes in Computer Science*, pages 108–118. Springer Berlin Heidelberg, 2008.

[3] Lacy G McNamee, Brittany L Peterson, and Jorge Peña. A call to educate, participate, invoke and indict: Understanding the communication of online hate groups. *Communication Monographs*, 77(2):257–280, 2010.

[4] S. Rawat and D.R. Patil. Efficient focused crawling based on best first search. In *Advance Computing Conference (IACC), 2013 IEEE 3rd International*, pages 908–911, Feb 2013.

[5] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, and Sidharth Chhabra. Mining youtube to discover extremist videos, users and hidden communities. In Pu-Jen Cheng, Min-Yen Kan, Wai Lam, and Preslav Nakov, editors, *Information Retrieval Technology*, volume 6458 of *Lecture Notes in Computer Science*, pages 13–24. Springer Berlin Heidelberg, 2010.

[6] Yilu Zhou, Edna Reid, Jialun Qin, Hsinchun Chen, and Guanpi Lai. Us domestic extremist groups on the web: link and content analysis. *Intelligent Systems, IEEE*, 20(5):44–51, 2005.