Discovery and retrieval of terrorism related content on the Web and social media

Dr. Theodora Tsikrika

Multimedia Knowledge and Social Media Analytics Lab
Information Technologies Institute
Centre for Research and Technology Hellas (CERTH)
HomeMade Explosives (HMEs) and Recipes characterisation

FP7 project (Nov 2013 - Dec 2016)

https://www.homer-project.eu/

reTriEval and aNalysis of heterogeneous online content for terrOrist activity Recognition

H2020 IA (Sep 2016 – Aug 2019)

http://tensor-project.eu/
The Web

- World Wide Web
  - Surface Web
    - Accessible, Indexed
  - Deep Web
    - Accessible, Non-indexed
    - Dark Web
      - Restricted, non-indexed
Surface vs. Deep vs. Dark Web

- **Surface Web:**
  - Readily accessible content
  - Indexed by search engines

- **Deep Web:**
  - Further user actions needed in order to access the content
  - Special techniques needed for crawling/indexing
  - Much larger than Surface Web

- **Dark Web:**
  - Special software needed in order to access the content
  - Provides users with anonymity
  - Includes several darknets (e.g., TOR, I2P, Freenet, etc.)
  - Usage: illegal marketplaces, whistleblowing, Bitcoin transactions, etc.
  - User base: from journalists and LEAs to criminals
Motivation

• **Challenges** for Law Enforcement Agencies (LEAs):
  – Extensive use of Surface Web & Dark Web for communication and diffusion of terrorism-related information
    • Propaganda and radicalization
    • Tutorials on the construction of explosives and weapons
  – Need for effective and efficient domain-specific discovery tools

• **Barriers:**
  – Surface Web discovery tools:
    • effective for general search, more limited for domain-specific search
  – Dark Web discovery tools:
    • limited for both general & domain-specific search
Domain-specific discovery methods

1. Focused crawling
   - Domain-specific document collection
   - Automatically traversing the Web link structure of the Web
   - Selecting links to follow based on their relevance to the domain

2. Search engine querying
   - Automatically query search engines/social media using their APIs
   - (Semi-)automatic domain-specific query generation & expansion

3. Hybrid approach
   - (1) + (2) + (post-retrieval classification)
Crawling

- Seed URLs
- frontier
- Fetch page
- Parse page
- discovered
- Extract links

- [Home](http://www.url1.com)
- [Picric acid](http://www.url2.com)
- [Nitroglycerine](http://www.url3.com)

Check out my website on my other hobby: [motorcycles](http://www.url4.com)
Focused crawling

1. Seed URLs
2. Frontier
3. Fetch page
4. Parse page
5. Represent link
6. Textual features
7. Classifier
8. Score > threshold

- [a href="http://www.url1.com"] Home
- [a href="http://www.url2.com"] Picric acid
- [a href="http://www.url3.com"] Nitroglycerine
- Check out my web site on my other hobby: [a href="http://www.url4.com"] motorcycles
Focused crawling

- Classifier-guided link selection
  - Anchor text
  - URL terms
  - Text window (x = 100 characters) surrounding anchor text
  - Web page text

Seed URLs

frontier

Fetch page

Parse page

Represent link

Classifier

score > threshold

<a href="http://www.url1.com">Home</a>

<a href="http://www.url2.com">Picric acid</a>

<a href="http://www.url3.com">Nitroglycerine</a>

Check out my web site on my other hobby: <a href="http://www.url4.com">motorcycles</a>
Focused crawling (+ Dark Web)

Seed URLs → Frontier

Fetch page
- Surface Web Module
- Tor Module
- I2P Module
- Freenet Module

Parse page → Represent link → Textual features → Classifier

score > threshold
Experiments

- Seed set: 5 pages (1 Surface Web, 1 TOR, 2 I2P, 1 Freenet)
- Seed set obtained: LEAs representatives + domain experts
- Crawling depth = 2
- Link selection classifier / Web page classifier
  - Training set: 400 (105 pos, 295 neg) / 600 (250 pos, 350 neg) samples
  - SVM classifier with RBF kernel

<table>
<thead>
<tr>
<th></th>
<th>Threshold</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Link-based classifier</strong></td>
<td>Precision</td>
<td>0.63</td>
<td>0.63</td>
<td>0.77</td>
<td>0.77</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>1.00</td>
<td>0.91</td>
<td>0.87</td>
<td>0.84</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>F-measure</td>
<td>0.77</td>
<td>0.74</td>
<td><strong>0.82</strong></td>
<td>0.8</td>
<td>0.58</td>
</tr>
<tr>
<td><strong>Link-based classifier</strong> + <strong>Web page classifier</strong></td>
<td>Precision</td>
<td>0.86</td>
<td>0.87</td>
<td>0.87</td>
<td>0.87</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Recall</td>
<td>1.00</td>
<td>0.99</td>
<td>0.96</td>
<td>0.92</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>F-measure</td>
<td><strong>0.93</strong></td>
<td>0.92</td>
<td>0.91</td>
<td>0.9</td>
<td>0.62</td>
</tr>
</tbody>
</table>
Search engine querying

• **Query generation & expansion**
  1. Exploit domain-specific knowledge for query generation
  2. Apply machine learning/deep learning for query expansion

• **Query submission**
  1. Multiple queries automatically submitted
  2. Search results merged (duplicate removal, re-ranking)
  3. Post-retrieval classification (filtering step)
## Query generation - patterns

<table>
<thead>
<tr>
<th>Concepts</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>explosive</em></td>
<td>acetone peroxide, anfo, c-4, hmtd, lead azide, lead picrate, mercury fulminate, nitrocellulose, nitrogen triiodide, nitroglycerin, nitroglycol, potassium chlorate, petn, picric acid, rdx, r-salt, semtex, tatp, trinitrotoluene TNT, urea nitrate</td>
</tr>
<tr>
<td><em>ingredient</em></td>
<td>ammonium nitrate, potassium nitrate</td>
</tr>
<tr>
<td><em>context</em></td>
<td>anarchist, islam</td>
</tr>
<tr>
<td><em>object</em></td>
<td>bomb(s), explosive(s), ied, pyrotechnics, homemade bomb(s), homemade explosive(s), homemade ied, homemade pyrotechnics, improvised bomb(s), improvised explosive(s), improvised pyrotechnics</td>
</tr>
<tr>
<td><em>action</em></td>
<td>how to make, manufacture, making, preparation, synthesis</td>
</tr>
<tr>
<td><em>recipe</em></td>
<td>recipe(s), preparatory manual</td>
</tr>
<tr>
<td><em>resource</em></td>
<td>book, forum, handbook, pdf, torrent, video</td>
</tr>
<tr>
<td>Patterns</td>
<td>Equivalent</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
</tr>
<tr>
<td><em>ingredient</em></td>
<td></td>
</tr>
<tr>
<td><em>explosive</em></td>
<td></td>
</tr>
<tr>
<td><em>explosive</em> <em>object</em></td>
<td></td>
</tr>
<tr>
<td><em>explosive</em> plastic</td>
<td></td>
</tr>
<tr>
<td>homemade <em>explosive</em></td>
<td></td>
</tr>
<tr>
<td><em>object</em></td>
<td></td>
</tr>
<tr>
<td><em>object</em> <em>recipe</em></td>
<td><em>recipe</em> <em>object</em></td>
</tr>
<tr>
<td><em>action</em> <em>explosive</em></td>
<td><em>explosive</em> <em>action</em></td>
</tr>
<tr>
<td><em>action</em> <em>explosive</em> at home</td>
<td><em>explosive</em> <em>action</em> <em>object</em></td>
</tr>
<tr>
<td><em>action</em> <em>explosive</em> <em>object</em></td>
<td><em>explosive</em> <em>object</em> <em>action</em> <em>explosive</em></td>
</tr>
<tr>
<td><em>action</em> <em>explosive</em> powder</td>
<td></td>
</tr>
<tr>
<td><em>action</em> <em>object</em></td>
<td><em>object</em> <em>action</em></td>
</tr>
<tr>
<td><em>action</em> <em>action</em> <em>explosive</em></td>
<td></td>
</tr>
</tbody>
</table>
## Query generation - patterns

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Candidate Queries</th>
</tr>
</thead>
</table>
| `_object_ _recipe_` | homemade bomb recipe  \  
homemade explosive recipe  \  
improvised bomb recipe  \  
improvised explosive recipe  \  
ied recipe           |
Query generation - patterns

Experimental evaluation

- 414 queries
- top 10 results retrieved
- 1157 unique URLs
- manually assessed
Query expansion

• Machine learning techniques (decision trees) for generating candidate expansion terms

problem OR bombs OR home OR time OR impact OR ^glass OR heating OR terms OR acid OR ^power OR ^rights OR ^time OR grams OR alcohol OR cap OR fuel OR reaction
OR (explosive AND ^petn)
OR (explosive AND ^world)
OR (explosive AND acid)

• Simplification

heating OR grams OR fuel OR reaction OR (explosive AND acid)
Hybrid discovery approach
Social media discovery framework
Key player identification

- **Aim:** identify key players in terrorism-related social media networks

- **Goal:** remove key players ➔ destroy internal connectivity ➔ community becomes small isolated networks

- **Motivation:** social media networks exhibit scale free topology
  - power law degree distribution
  - robust to random attacks
  - vulnerable to targeted attacks

- **Approach:** targeted attacks based on centrality measures (existing+new)

- **Evaluation:** social media network of terrorism-related Twitter posts
Terrorism-related social media discovery

• Social media network of Twitter accounts
  – query Twitter API
  – Arabic keywords provided by LEAs + domain experts
  – keywords related to Caliphate state (ISIS)

• Dataset:
  – 38,766 posts by 5,461 users
  – 100 posts manually assessed for relevance
  – users linked through mentions
  – largest connected component: 3,600 users/9,203 links
  – 2.56 power law exponent (p-value = 0.7780)
Results: largest connected component decay

Decrease in relative size:
- 5% random attack
- 27.1% closeness centrality
- 44 – 49% rest of centrality measures
- 50.1% MEB
Results: key players

• Top-10 key players identified by each of the 7 centrality measures
  – 18 unique Twitter user accounts

• 10 days after dataset construction:
  – 14 out of 18 suspended
  – 10 out of 14 suspensions took place within 72 hours of account creation

• Further evidence to dataset relevance

• High volatility
Conclusions

• Domain-specific discovery tools
  – Build your own search engine
  – Exploit capabilities of already existing search systems
  – Combine them in a hybrid approach
  – Exploit social network structures

• Challenges
  – Multilingual and Multimedia content
  – From Surface to Dark Web
  – Volatility (Dark Web, social media)
  – Validating sources (mis-information, dis-information, etc.)
  – Legal, ethical and privacy aspects
References


4. C. Iliou, G. Kalpakis, T. Tsikrika, S. Vrochidis, I. Kompatsiaris, "Hybrid Focused Crawling for Homemade Explosives Discovery on Surface and Dark Web", 11th International Conference on Availability, Reliability and Security (ARES 2016), Salzburg, Austria, Aug 2016


THANK YOU!

http://mklab.iti.gr

theodora.tsikrika@iti.gr