Filtering of Multi-Lingual Terrorist Content with Graph-Theoretic Classification Tools

Mark Last
Ben-Gurion University of the Negev, Beer-Sheva, Israel

In cooperation with
Abraham Kandel (USF), Alex Markov (BGU), Dror Magal (Meged)

An up-to-date version of this tutorial is available at
Outline

• Introduction
  – Internet as a Terrorist Weapon
  – Selected Examples of Multi-Lingual Terrorist Content
  – Challenges in Filtering Terrorist Content

• Web Document Representation and Categorization
  – The Vector-Space Approach
  – The Graph-Based Approach
  – The Hybrid Approach

• Case Studies

• Conclusions and Future Work
Important Assumptions

• The terrorist organizations mentioned in this tutorial are included in the list of U.S.-Designated Foreign Terrorist Organizations, which is updated periodically by the U.S. Department of State, Office of Counterterrorism.
  – The latest list can be downloaded from http://www.infoplease.com/ipa/A0908746.html

• Affiliations of specific web sites with terrorist organizations are available from several sources such as:
  – SITE Institute http://www.siteinstitute.org/
  – Internet Haganah http://www.haganah.org.il/
  – The Intelligence and Terrorism Information Center http://www.terrorism-info.org.il
Internet as a Terrorist Weapon

- A full range of instructions for terrorist attacks, including maps, photographs, directions, codes and even technical details of how to use the bombs are being transferred through the Internet, Cyber-terrorism, *Foreign Report*, London, 1997

- The Internet's largest threat is simply the ease of international communication and the ability to hide among the seemingly infinite volume of traffic it carries, *Robert Lemos, ZDNet, August 26, 2002*

- "They lost their base in Afghanistan, they lost their training camps, they lost a government that allowed them do what they want within a country. Now they're surviving on internet to a large degree. It is really their new base“, *Peter Bergen, October 6, 2004*
What information is posted by terrorists?

- Propaganda (for insiders and outsiders)
- Fundraising solicitations
- Basic training
  - How to mix ricin poison, how to make a bomb from commercial chemicals, how to sneak through Syria into Iraq, etc.
  - A country-by-country list of "explosive materials available in Western markets"
- Specific orders
  - Madrid – March 2004
    - "[The Islamist cell] took its inspiration from a Web site that called on local Islamists to stage attacks in Spain before the 2004 general elections to prompt withdrawal of troops from Iraq", [the court spokeswoman] said. (The New York Times, April 11, 2006)
  - London – July 2005
    - A message posted on May 29 on an Islamist Internet site: "We ask all waiting mujahedeen, wherever they are, to carry out the planned attack" (The New York Times, July 13, 2005)
    - "The July 7 bombings in London were a low-budget operation carried out by four men who had no connection to Al Qaeda and who obtained all the information they needed from the Internet" (The New York Times, April 11, 2006)
Terrorist Content

Selected Examples
Filtering of Multi-Lingual Terrorist Content

Sabiroon - Hamas
Language: English
Une nouvelle unité armée israélienne a repoussé les manifestations pacifiques palestiniennes

Agences
Des sources israéliennes ont livré le passage suivant: pour la première fois, sur la constitution d'une nouvelle unité militaire israélienne par l'armée de l'occupation pour disperser une manifestation organisée par les Palestiniens dans le village de Safin.

Mise à jour: 8 mai 2006, 19h00

Actualités
Les lycéens occupent l'eau des maisons civiles dans le quartier de Tel Al-Roumata
Al-Khalil - Spécial
Plusieurs groupes du quartier de Tel Al-Roumata, au milieu de la ville d'Al-Khalil, affirment que les colons israéliens ont saisi plusieurs centaines d'eau et court l'eau de leurs maisons, dans l'objectif de les forcer à quitter leur quartier.

Mise à jour: 8 mai 2006, 19h00

Actualités
Ariel Sharon appelle à l'immobilisation de toute nouvelle libération de prisonniers palestiniens
JERUSALEM (AP)
Le Premier ministre israélien Ariel Sharon s'est ému dimanche de toute nouvelle libération de prisonniers palestiniens, tôt que l'Autorité de Mahmoud Abbas n'a pas pris des mesures plus répressives contre les groupes radicaux.

Mise à jour: 9 mai 2006, 19h00

Actualités
Rapport: Lesапример réfutant des voix en Cisjordanie et dans la bande de Gaza
Agences
Les résultats déclaraient que la participation de la resistance islamique du Hamas et de mouvement de la libération nationale du Fatah sont contradictoires.

Mise à jour: 9 mai 2006, 19h00
Filtering of Multi-Lingual Terrorist Content

Qudsway – Palestinian Islamic Jihad
Language: Arabic
Army of Ansar Al-Sunna (Iraq)
Language: Arabic
Mark Last (BGU)

Filtering of Multi-Lingual Terrorist Content

Hezbollah (Lebanon)

Language: Hebrew
Challenges in Filtering Terrorist Content

• **Finding relevant content in multiple languages**
  – Terrorist web sites frequently switch their URLs
  – There is more online information *about* terrorists than information created and posted *by* terrorists
  – What makes terrorist content different from a regular news report or commentary?

• **Terrorist group identification**
  – The true web site affiliation is often concealed
    • How can we tell that the “Palestinian Information Center” is associated with Hamas?

• **Topic identification**
  – Propaganda, fundraising, bomb-making, etc.

• **Real-time understanding of multi-lingual content**
  – On Sept. 10, 2001, the NSA intercepted two Arabic-language messages, “Tomorrow is zero hour” and "The match is about to begin." The sentences weren't translated until Sept. 12, 2001 (Michael Erard, MIT Technology Review, March 2004)
Text Categorization (TC)  
Basic Definition

• TC – task of assigning a Boolean \{T, F\} value to each pair \(\langle d_i, c_i \rangle \in D \times C\)

  where
  \(D = (d_1, \ldots, d_{|D|})\) is a collection of documents
  \(C = (c_1, \ldots, c_{|C|})\) is a set of pre-defined categories

  –Sample categories: “terrorist”, “non-terrorist”, “bomb-making”, etc.
Inductive text classification / categorization

• The Goal
  – Infer a classification model from a representative sample of labeled training documents

• Requirements in the Terrorist Domain
  – High accuracy
    • The correct category/ categories of each document should be identified as accurately as possible
  – Interpretability
    • An automatically induced model should be subject to scrutiny by a human expert
  – Speed
    • The model should be capable to process massive streams of web documents in minimal time
  – Multilinguality
    • The model induction methods should maintain a high performance level over web content in multiple languages
Text Categorization (TC) Tasks

- **Binary TC** – two non-overlapping categories only
  - Example: “terrorist” vs. “non-terrorist”
- **Multi-Class TC** – more than two non-overlapping categories
  - Example: “PIJ” or “Hamas” or “Al-Aqsa Brigades”
  - A multi-class problem can be reduced into multiple binary tasks (one-against-the-rest strategy)
- **Multi-Label TC** – overlapping categories are allowed
  - Example: a “Hamas” document on “bomb-making”
  - A multi-label task can be split into a set of binary classification tasks
- **Ranking categorization**
  - *Category ranking*: which categories match a given document best?
  - *Document ranking*: which documents match a given category best?
The Vector-Space Model
(Salton et al., 1975)

- A text document is considered a “bag of words (terms / features)”
  - Document \( d_j = (w_{1j}, \ldots, w_{|T|j}) \) where \( T = (t_1, \ldots, t_{|T|}) \) is set of terms (features) that occurs at least once in at least one document (vocabulary)
- Term: \( n \)-gram, single word, noun phrase, keyphrase, etc.
- Term weights: binary, frequency-based, etc.
- Meaningless (“stop”) words are removed
- Stemming operations may be applied
  - Leaders => Leader
  - Expiring => expire
- The ordering and position of words, as well as document logical structure and layout, are completely ignored
Term Weighting
(Salton and McGill, 1983)

- **Binary**
  \[ w_{ij} = \begin{cases} 
  1, & \text{if a term } t_j \text{ occurs in document } d_i \\
  0, & \text{otherwise} 
  \end{cases} \]

- **Normalized Term Frequency**
  \[ w_{ij} = \frac{TF_{ij}}{\max_j TF_{ij}} \]
  where \( TF_{ij} = \text{raw frequency of term } t_j \text{ in document } d_i \)

- **TFIDF (term frequency × inverse document frequency)**
  \[ w_{ij} = TF \times IDF = TF_{ij} \times \log \frac{N}{n} \]
  where
  \[ N = \text{number of documents in collection (corpus)} \]
  \[ n = \text{number of documents where term } t_j \text{ occurs at least once} \]
Earlier, Khaled Mishaal, the Movement's top political leader, said in a rally in the Palestinian refugee camp of Yarmouk in the Syrian capital, Damascus, Friday that there was no more room for further calm in the light of the Israeli daily hostilities against the Palestinian people.

By ASSOCIATED PRESS
Dec. 10, 2005
Hamas will not renew its truce with Israel when it expires at the end of the year, the political leader of the Palestinian terrorist group, Khaled Mashaal, told a rally Friday.

Expires Friday group Hamas Israel Khaled leader Mashaal Palestinian political rally renew terrorist truce year
The “Bag of Words” Approach
A Practical Example

Bag of Words 1

Friday further hostilities Israel Khaled leader light Mishaal Movement Palestinian people political rally refugee room Syrian top Yarmouk

8 words in common!

Bag of Words 2

Expires Friday group Hamas Israel Khaled leader Mashaal Palestinian political rally renew terrorist truce year

Terrorist

Non-Terrorist
Automated Keyphrase Extraction
(Turney, 2000)

- Term definition
  - *Keyphrase* = a sequence of one, two, or three words that appear consecutively in the text, with no intervening stop words or punctuation marks
  - Example: “Palestinian Islamic Jihad”

- Keyphrase weight
  - Phrase frequency in the text multiplied by a factor

- The maximum number of keyphrases in a document is a user-specified parameter (default = 10)

- The best phrase classification model is found by a genetic algorithm
  - The model has been induced from corpora in English
  - The model is proprietary
  - Estimated processing speed: 2k – 3k HTML documents per second on a Pentium III processor
Advantages of the Vector-Space Model
(based on Joachims, 2002)

- A simple and straightforward representation for English and other languages, where words have a clear delimiter
- Most weighting schemes require a single scan of each document
- A fixed-size vector representation makes unstructured text accessible to most classification algorithms (from decision trees to SVMs)
- Consistently good results in the information retrieval domain (mainly, on English corpora)
Limitations of the Vector-Space Model

- **Text documents**
  - Ignoring the *word position* in the document
  - Ignoring the *ordering of words* in the document

- **Web Documents**
  - Ignoring the information contained in HTML tags (e.g., document sections)

- **Multilingual documents**
  - Word separation may be tricky in some languages (e.g., Latin, German, Chinese, etc.)
  - No comprehensive evaluation on large non-English corpora
DIVIDE ET IMPERA
(“Divide and Rule”)
The Word Separation in the Ancient Latin

The Arch of Titus, Rome
(1st Century AD)

Dedication to Julius Caesar
(1st Century BC)

Words are separated by triangles
Alternative Representation of Multilingual Web Documents:

The Graph-Based Model

(introduced in Schenker et al., 2005)
Relevant Definitions
(Based on Bunke and Kandel, 2000)

- A (labeled) graph $G$ is a 4-tuple $G = (V, E, \alpha, \beta)$

Where

$V$ is a set of nodes (vertices), $E \subseteq V \times V$ is a set of edges connecting the nodes, $\alpha$ is a function labeling the nodes and $\beta$ is a function labeling the edges.

- Node and edge IDs are omitted for brevity

**Graph size:** $|G| = |V| + |E|$
The Graph-Based Model of Web Documents

- **Basic ideas:**
  - one node for each unique term
  - if word $B$ follows word $A$, there is an edge from $A$ to $B$
    - In the presence of terminating punctuation marks (periods, question marks, and exclamation points) no edge is created between two words
  - stop words are removed
  - graph size is limited by including only the most frequent terms
  - **Stemming**
    - Alternate forms of the same term (singular/plural, past/present/future tense, etc.) are conflated to the most frequently occurring form
  - Several variations for node and edge labeling (see the next slides)
The *Standard* Representation

- Edges are labeled according to the document section where the words are followed by each other
  - *Title (TI)* contains the text related to the document’s title and any provided keywords (meta-data);
  - *Link (L)* is the “anchor text” that appears in clickable hyper-links on the document;
  - *Text (TX)* comprises any of the visible text in the document (this includes anchor text but not title and keyword text)
The *Simple* Representation

- The graph is based only the visible text on the page (title and meta-data are ignored)
- Edges are not labeled
The \textit{n-distance} Representation

- Based on the visible text only
- Instead of considering only terms immediately following a given term in a web document, we look up to \( n \) terms ahead and connect the succeeding terms with an edge that is labeled with the distance between them (unless the words are separated by certain punctuation marks)
- \( n \) is a user-provided parameter.

\( n = 3 \)
The \( n\)-simple Representation

- Based on the visible text only
- We look up to \( n \) terms ahead and connect the succeeding terms with an unlabeled edge
- \( n \) is a user-provided parameter.

\( n = 2 \)

\( n = 3 \)
The Absolute Frequency Representation

- No section-related information
- Each node and edge is labeled with an absolute frequency measure
The Relative Frequency Representation

- No section-related information
- Each node and edge is labeled with a relative frequency measure
- A normalized value in [0,1] is assigned by dividing each node frequency value by the maximum node frequency value that occurs in the graph
- A similar procedure is performed for the edges
Iraq bomb: Four dead, 110 wounded

A car bomb has exploded outside a popular Baghdad restaurant, killing three Iraqis and wounding more than 110 others, police officials said. Earlier an aide to the office of Iraqi Prime Minister Ibrahim al-Jaafari and his driver were killed in a drive-by shooting.

FULL STORY
Graph Based Document Representation - Parsing

<title>CNN.com International</title>


<h2>Iraq bomb: Four dead, 110 wounded</h2>

A car bomb has exploded outside a popular Baghdad restaurant, killing three Iraqis and wounding more than 110 others, police officials said. Earlier an aide to the office of Iraqi Prime Minister Ibrahim al-Jaafari and his driver were killed in a drive-by shooting.

A car bomb exploded outside a popular Baghdad restaurant, killing three Iraqis and wounding more than 110 others, police officials said. Earlier an aide to the office of Iraqi Prime Minister Ibrahim al-Jaafari and his driver were killing in a drive-by shooting.

Links
Iraq bomb: Four dead, 110 wounded. FULL STORY.
A car bomb has exploded outside a popular Baghdad restaurant, killing \textit{three Iraqis} and wounding more than 110 others, police officials said. Earlier an aide to the office of Iraqis Prime Minister Ibrahim al-Jaafari and his driver were \textit{killing} in a driver shooting.

\textbf{Links}

Iraqis bomb: Four dead, 110 wounding.
FULL STORY.
### Ten most frequent terms are used

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iraqis</td>
<td>3</td>
</tr>
<tr>
<td>Killing</td>
<td>2</td>
</tr>
<tr>
<td>Bomb</td>
<td>2</td>
</tr>
<tr>
<td>Wounding</td>
<td>2</td>
</tr>
<tr>
<td>Driver</td>
<td>2</td>
</tr>
<tr>
<td>Exploded</td>
<td>1</td>
</tr>
<tr>
<td>Baghdad</td>
<td>1</td>
</tr>
<tr>
<td>International</td>
<td>1</td>
</tr>
<tr>
<td>CNN</td>
<td>1</td>
</tr>
<tr>
<td>Car</td>
<td>1</td>
</tr>
</tbody>
</table>
Simple Graph Based Document Representation

Ten most frequent terms are used

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iraqis</td>
<td>3</td>
</tr>
<tr>
<td>Killing</td>
<td>2</td>
</tr>
<tr>
<td>Bomb</td>
<td>2</td>
</tr>
<tr>
<td>Wounding</td>
<td>2</td>
</tr>
<tr>
<td>Driver</td>
<td>2</td>
</tr>
<tr>
<td>Exploded</td>
<td>1</td>
</tr>
<tr>
<td>Baghdad</td>
<td>1</td>
</tr>
<tr>
<td>International</td>
<td>1</td>
</tr>
<tr>
<td>CNN</td>
<td>1</td>
</tr>
<tr>
<td>Car</td>
<td>1</td>
</tr>
</tbody>
</table>
“Lazy” Categorization with Graph-Based Models

• The Basic $k$-Nearest Neighbors Algorithm
  – Input: a set of labeled training documents, a query document $d$, and a parameter $k$ defining the number of nearest neighbors to use
  – Output: a label indicating the category of the query document $d$
  – Step 1. Find the $k$ nearest training documents to $d$ according to a distance measure
  – Step 2. Select the category of $d$ to be the category held by the majority of the $k$ nearest training documents

• $k$-Nearest Neighbors with Graphs (Schenker et al., 2005)
  – Represent the documents as graphs (done)
  – Use a graph-theoretical distance measure
Distance between two Graphs

• Required properties
  – (1) boundary condition: \( d(G_1, G_2) \geq 0 \)
  – (2) identical graphs have zero distance:
    \[ d(G_1, G_2) = 0 \rightarrow G_1 \cong G_2 \]
  – (3) symmetry: \( d(G_1, G_2) = d(G_2, G_1) \)
  – (4) triangle inequality:
    \[ d(G_1, G_3) \leq d(G_1, G_2) + d(G_2, G_3) \]
Relevant Definitions

(Based on Bunke and Kandel, PRL, 2000)

• A graph \(G_1 = (V_1, E_1, \alpha_1, \beta_1)\) is a **sub-graph** of a graph \(G_2 = (V_2, E_2, \alpha_2, \beta_2)\), denoted \(G_1 \subseteq G_2\), if \(V_1 \subseteq V_2\), \(E_1 \subseteq E_2 \cap (V_1 \times V_1)\), \(\alpha_1(x) = \alpha_2(x) \ \forall x \in V_1\) and \(\beta_1(x, y) = \beta_2(x, y) \ \forall (x, y) \in E_1\).

• Conversely, the graph \(G_2\) is also called a **supergraph** of \(G_1\).

\[
\begin{align*}
A & \xrightarrow{x} B \\
\end{align*}
\]

\[
\begin{align*}
A & \xrightarrow{x} B \quad \text{(\(G_1\))} \\
\end{align*}
\]

\[
\begin{align*}
A & \xrightarrow{x} B \xrightarrow{y} C \\
\end{align*}
\]

\[
\begin{align*}
A & \xrightarrow{x} B \quad \text{(\(G_2\))} \\
\end{align*}
\]
More Graph-Theoretic Definitions

- A graph $G_1 = (V_1, E_1, \alpha_1, \beta_1)$ and a graph $G_2 = (V_2, E_2, \alpha_2, \beta_2)$ said to be isomorphic, denoted $G_1 \cong G_2$, if there exists a bijective function $f : V_1 \to V_2$ such that $\alpha_1(x) = \alpha_2(f(x)) \forall x \in V_1$ and $\beta_1(x, y) = \beta_2(f(x), f(y)) \forall (x, y) \in V_1 \times V_1$. 

\[
\begin{array}{c}
A & \xrightarrow{x} & B \\
\downarrow{w} & & \downarrow{z} \\
C & & D \\
\end{array}
\hspace{1cm}
\begin{array}{c}
A & \xrightarrow{x} & B \\
\downarrow{w} & & \downarrow{z} \\
B & & C \\
\end{array}
\]
More Graph-Theoretic Definitions

- **Subgraph Isomorphism** – graph is isomorphic to a part (subgraph) of another graph
- **Graph isomorphism** is not known as NP-complete
- **Subgraph isomorphism** is NP-complete.
More Graph-Theoretic Definitions

- Let $G$, $G_1$ and $G_2$ be graphs. The graph $G$ is a **common subgraph** of $G_1$ and $G_2$ if there exist subgraph isomorphisms from $G$ to $G_1$ and from $G$ to $G_2$. 

$$
\begin{align*}
\text{Graph } G_1 & \quad \text{Graph } G \\
A & \rightarrow B \quad A \rightarrow B \\
C & \rightarrow D & B & \rightarrow E \\
w & \rightarrow z & x & \rightarrow x \\
& \rightarrow y & & y & \rightarrow r \\
\end{align*}
$$

$$
\begin{align*}
\text{Graph } G_2 & \\
A & \rightarrow B \rightarrow F \\
x & \rightarrow \quad p & \rightarrow r \\
\end{align*}
$$
More Graph-Theoretic Definitions (cont.)

- The graph $G$ is a **maximum common subgraph (mcs)** if $G$ is a common subgraph of $G_1$ and $G_2$ and there exist no other common subgraph $G'$ of $G_1$ and $G_2$ such that $|G'| > |G|$

\[
|G| = |V| + |E| = 2 + 1 = 3
\]
More Graph-Theoretic Definitions (cont.)

- Let $G$, $G_1$ and $G_2$ be graphs. The graph $G$ is a **common supergraph** of $G_1$ and $G_2$ if there exist subgraph isomorphisms from $G_1$ to $G$ and from $G_2$ to $G$.
More Graph-Theoretic Definitions (cont.)

- The graph $G$ is a **minimum common supergraph (MCS)** if $G$ is a common supergraph of $G_1$ and $G_2$ and there exist no other common supergraph $G'$ of $G_1$ and $G_2$ such that $|G'| < |G|$

\[ |G| = |V| + |E| = 4 + 2 = 6 \]
Distance between two Graphs

- MMCSN Measure (Schenker et al., 2005):

\[ d_{MMCSN}(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{|MCS(G_1, G_2)|} \]

- \( mcs(G_1, G_2) \) - maximum common subgraph
- \( MCS(G_1, G_2) \) - minimum common common supergraph

\[ d_{MMCSN}(G_1, G_2) = 1 - \frac{2 + 1}{4 + 5} = 0.667 \]
Other Distance Measures

- Bunke and Shearer (1998): \[ d_{MCS}(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{\max(|G_1|, |G_2|)} \]

- Wallis et al. (2001): \[ d_{WGU}(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{|G_1| + |G_2| - |mcs(G_1, G_2)|} \]

- Bunke (1997): \[ d_{UGU}(G_1, G_2) = |G_1| + |G_2| - 2|mcs(G_1, G_2)| \]

- Fernández and Valiente (2001): \[ d_{MMCS}(G_1, G_2) = |MCS(G_1, G_2)| - |mcs(G_1, G_2)| \]
k-Nearest Neighbors with Graphs
Empirical Evaluation

- Benchmark Data Set: K-series
  - Source: Boley et al., 1999
  - 2,340 web documents from 20 categories
  - Documents in this collection were originally English news pages hosted at Yahoo!
  - The data set is available at:
  - List of news categories:
    - business, health, politics, sports, technology, entertainment, art, cable, culture, film, industry, media, multimedia, music, online, people, review, stage, television, and variety
k-Nearest Neighbors with Graphs

Accuracy vs. Graph Size

- Vector model (cosine)
- Vector model (Jaccard)
- Graphs (40 nodes/graph)
- Graphs (70 nodes/graph)
- Graphs (100 nodes/graph)
- Graphs (150 nodes/graph)
k-Nearest Neighbors with Graphs

Accuracy vs. Distance Measure
k-Nearest Neighbors with Graphs
Accuracy vs. Graph Representation
### k-Nearest Neighbors with Graphs

**Average Time to Classify One Document**

<table>
<thead>
<tr>
<th>Method</th>
<th>Average time to classify one document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector (cosine)</td>
<td>7.8 seconds</td>
</tr>
<tr>
<td>Vector (Jaccard)</td>
<td>7.79 seconds</td>
</tr>
<tr>
<td>Graphs, 40 nodes/graph</td>
<td>8.71 seconds</td>
</tr>
<tr>
<td>Graphs, 70 nodes/graph</td>
<td>16.31 seconds</td>
</tr>
<tr>
<td>Graphs, 100 nodes/graph</td>
<td>24.62 seconds</td>
</tr>
</tbody>
</table>
**k-Nearest Neighbors with Graphs**

- **Advantages**
  - Keeps HTML structure information
  - Retains original order of words
  - More accurate than $k$-NN with the vector-space model

- **Limitation**
  - Very low classification speed
    - Up to three times slower than vector classification

- **Conclusion**
  - Graph models cannot be used for real-time filtering of web documents
The Hybrid Approach to Document Categorization
(Markov et al., 2006)

• Basic Idea
  – Represent a document as a vector of sub-graphs
  – Categorize documents with a model-based classifier (e.g., a decision tree), which is much faster than a “lazy” method

• Naïve Approach
  – Select sub-graphs that are most frequent in each category

• Smart Approach
  – Select sub-graphs that are frequent in a specific category and not frequent in other categories
Predictive Model Induction with Hybrid Representation

Set of documents with known categories - the training set

Documents graph representation

Extraction of sub-graphs relevant for classification

Graph Construction

Web or text documents

Subgraph Extraction

Text representation

Creation of prediction model

Feature selection (optional)

Document classification rules

Representation of all documents as vectors with Boolean values for every sub-graph in the set

Identification of best attributes (boolean features) for classification

Finally - prediction model induction and extraction of classification rules
Subgraph Extraction – The Naïve Approach

• Input:
  – $G$ – A training set of document graphs
  – $t_{\text{min}}$ – Threshold (minimum subgraph frequency)

• Output:
  – A set of classification-relevant subgraphs

• Process:
  – For each category, find frequent subgraphs $SCF > t_{\text{min}}$
  – $SCF$ (Subgraph Class Frequency): percentage of documents containing a subgraph in a given category
  – Combine all frequent subgraphs into one set

• Basic Assumption
  – Classification-Relevant Sub-Graphs are frequent in a specific category
Subgraph Extraction – The Smart Approach

- **Input**
  - \( G \) – training set of directed, unique nodes graphs
  - \( CR_{min} \) - Minimum Classification Rate

- **Output**
  - Set of classification-relevant sub-graphs

- **Process:**
  - For each class find subgraphs \( CR > CR_{min} \)
  - Combine all sub-graphs into one set

- **Basic Assumption**
  - **Classification-Relevant Sub-Graphs** are more frequent in a specific category than in other categories
The Smart Subgraph Extraction

• **SCF (Subgraph Class Frequency):**

\[
SCF(g'_k(c_i)) = \frac{g'_{kf}(c_i)}{N(c_i)}
\]

- \(SCF(g'_k(c_i))\) - frequency of sub-graph \(g'_k\) in category \(C_i\)
- \(N(c_i)\) - Number of documents in category \(C_i\)
- \(g'_{kf}(c_i)\) - Number of documents containing \(g'_k\) in category \(C_i\)
The Smart Subgraph Extraction (cont.)

• Inverse Subgraph Frequency:

\[
ISF(g'_k(c_i)) = \begin{cases} 
\log_2 \left( \frac{\sum N(c_j)}{\sum g'_kf(c_j)} \right) & \text{if } \sum g'_kf(c_j) > 0 \\
\log_2 (2 \times \sum N(c_i)) & \text{if } \sum g'_kf(c_j) = 0 
\end{cases} \quad \forall c_j \in C; \ j \neq i
\]

\(ISF(g'_k(c_i))\) - Inverse frequency of sub-graph in all categories except \(C_i\)

\(N(c_j)\) - Number of documents in category \(C_j\)

\(g'_kf(c_j)\) - Number of documents containing \(g'_k\) in category \(C_j\)
The Smart Subgraph Extraction (cont.)

- **Subgraph Classification Rate:**

\[ CR(g'_k(c_i)) = SCF(g'_k(c_i)) \times ISF(g'_k(c_i)) \]

- **SCF (g'\_k(c_i))** - Subgraph Class Frequency of subgraph \( g'_k \) in category \( c_i \)
- **ISF (g'\_k(c_i))** - Inverse Subgraph Frequency of subgraph \( g'_k \) in category \( c_i \)
- **Classification Relevant Feature** is a feature that best explains a specific category, or frequent in this category more than in all others
Subgraph Extraction – The Smart Approach with Fixed Threshold

• Input
  - \( G \) – training set of directed, unique nodes graphs
  - \( t_{\text{min}} \) – Threshold (minimum subgraph frequency)
  - \( CR_{\text{min}} \) - Minimum Classification Rate

• Output
  - Set of classification-relevant subgraphs

• Process:
  - For each class find subgraphs \( SCF > t_{\text{min}} \) and \( CR > CR_{\text{min}} \)
  - Combine all subgraphs into one set

• Basic Assumption
  - Classification-Relevant SubGraphs are frequent in a specific category and not frequent in other categories
### Frequent Subgraph Extraction: Notations

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$</td>
<td>Set of document graphs</td>
</tr>
<tr>
<td>$t_{min}$</td>
<td>Subgraph frequency threshold</td>
</tr>
<tr>
<td>$K$</td>
<td>Number of edges in the graph</td>
</tr>
<tr>
<td>$G$</td>
<td>Single graph</td>
</tr>
<tr>
<td>$sg$</td>
<td>Single subgraph</td>
</tr>
<tr>
<td>$sg^k$</td>
<td>Subgraph with $k$ edges</td>
</tr>
<tr>
<td>$F^k$</td>
<td>Set of frequent subgraphs with $k$ edges</td>
</tr>
<tr>
<td>$E^k$</td>
<td>Set of extension subgraphs with $k$ edges</td>
</tr>
<tr>
<td>$C^k$</td>
<td>Set of candidate subgraphs with $k$ edges</td>
</tr>
</tbody>
</table>
Frequent Subgraphs Extraction: The Naïve Algorithm
(based on the FSG algorithm by Kuramochi and Karypis, 2004)

1: $F^0 \leftarrow$ Detect all frequent single node subgraphs (nodes) in $G$
2: $k \leftarrow 1$
3: While $F^{k-1} \neq \emptyset$ Do
4: For Each subgraph $sg^{k-1} \in F^{k-1}$ Do
5: For Each graph $g \in G$ Do
6: If $sg^{k-1}$ is subgraph of $g$ Then
7: $E^k \leftarrow$ Detect all possible $k$ edge extensions of $sg^{k-1}$ in $g$
8: For Each subgraph $sg^k \in E^k$ Do
9: If $sg^k$ already a member of $C^k$ Then
10: $\{sg^k \in C^k\}.Count++$
11: Else
12: $sg^k$.Count $\leftarrow 1$
13: $C^k \leftarrow sg^k$
14: $F^k \leftarrow \{sg^k \in C^k \mid sg^k$.Count $> t_{min} \times |G|\}$
15: $k++$
16: Return $F^1, F^2, \ldots F^{k-2}$
Frequent Subgraph Extraction: Complexity

Assumption
A labeled vertex is unique in each graph

Subgraph isomorphism
Isomorphism between graph \(G_1=(V_1,E_1,\alpha_1,\beta_1)\) and part of graph \(G_2=(V_2,E_2,\alpha_2,\beta_2)\) can be found by two simple actions:

1. Determine that \(V_1 \subseteq V_2\) - \(O(|V_1| \cdot |V_2|)\)
2. Determine that \(E_1 \subseteq E_2\) - \(O(|V_1|^2)\)

Total complexity:
\(O(|V_1| \cdot |V_2| + |V_1|^2) \leq O(|V_2|^2)\)

Graph isomorphism
Isomorphism between graphs \(G_1=(V_1,E_1,\alpha_1,\beta_1)\) and \(G_2=(V_2,E_2,\alpha_2,\beta_2)\) can be found by two simple actions:

1. Determine \(G_1 \subseteq G_2\) - \(O(|V_2|)\)
2. Determine \(G_2 \subseteq G_1\) - \(O(|V_2|)\)

Total complexity: \(O(|V_2|)\)
Frequent Subgraph Extraction Example

Subgraphs

Document Graph

Extensions

Arab

Bank

Arab

Politic

West

Arab

West

Arab

Politic

Arab

Politic

Arab

Politic

Arab
Comparative Evaluation

• **Benchmark Data Sets**
  - K-series (Source: Boley et al., 1999)
    • 2,340 documents and 20 categories
    • Documents in those collections were originally news pages hosted at Yahoo
  - U-series (Source: Craven et al., 1998)
    • 4167 documents taken from the computer science department of four different universities: Cornell, Texas, Washington, and Wisconsin
    • 7 major categories: course, faculty, students, project, staff, department and other

• **Dictionary construction**
  - $N$ most frequent words in each document were taken for vector / graph construction, that is, exactly the same words in each document were used for both the graph-based and the bag-of-words representations
Vocabulary Size as a Function of Frequent Terms Used

![Graph showing the relationship between vocabulary size and frequent terms used per document. The graph plots frequent terms used per document on the x-axis and vocabulary size on the y-axis. Two series are shown: K-Series and U-Series.](image-url)
Classification Results with C4.5– K series data set

Accuracy Comparison for C4.5, K-series

Classification Accuracy

Frequent Terms Used

- Bag-of-words
- Hybrid Naïve
- Hybrid Smart
- Hybrid Smart with Fixed Threshold

Accuracy:
- 65%
- 70%
- 75%
- 80%
Classification Results with C4.5–U series data set

Accuracy Comparison for C4.5, U-series

Classification Speed: 0.3 sec. per 1,000 documents

Classification Speed: 1.7 sec. per 1,000 documents
Offline and Online Execution Times for C4.5

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Method</th>
<th>Time to Build Graphs (sec)</th>
<th>Time to Build Dictionary (sec)</th>
<th>Time to Construct Vectors (sec)</th>
<th>Time to Build Classification Model (sec)</th>
<th>Total Time Offline (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-series</td>
<td>Hybrid Smart ($N = 100$, $CR_{min} = 1.1$)</td>
<td>223.2</td>
<td>2628.56</td>
<td>5.59</td>
<td>4.36</td>
<td>2861.71</td>
</tr>
<tr>
<td></td>
<td>Hybrid Naïve ($N = 100$, $t_{min} = 0.1$)</td>
<td>223.2</td>
<td>43.4</td>
<td>31.16</td>
<td>76.59</td>
<td>374.35</td>
</tr>
<tr>
<td></td>
<td>Hybrid with Fixed Threshold ($N = 100$, $t_{min} = 0.1$, $CR_{min} = 0.1$)</td>
<td>223.2</td>
<td>66.35</td>
<td>7.47</td>
<td>6.09</td>
<td>303.11</td>
</tr>
<tr>
<td></td>
<td>Bag-of-words ($N = 20$)</td>
<td>n/a</td>
<td>300.9</td>
<td>133.2</td>
<td>330.32</td>
<td>764.42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Method</th>
<th>Average Time to Classify One Document (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-series</td>
<td>Hybrid Smart</td>
<td>$2.88 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>Hybrid Naïve</td>
<td>$4.56 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>Hybrid with Fixed Threshold</td>
<td>$3.12 \times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>Bag-of-words</td>
<td>$1.68 \times 10^{-3}$</td>
</tr>
</tbody>
</table>
Classification Results with Naïve Bayes – K series data set

Accuracy Comparison for NBC, K-series

Classification Accuracy

Frequent Terms Used

- Bag-of-words
- Hybrid Naïve
- Hybrid Smart
- Hybrid Smart with Fixed Threshold
Mark Last (BGU)

Classification Results with Naïve Bayes – U series data set

Accuracy Comparison for NBC, U-series

Classification Speed: 1.2 sec. per 1,000 documents

Classification Speed: 125 sec. per 1,000 documents

Accuracy Comparison for NBC, U-series

Classification Speed: 1.2 sec. per 1,000 documents

Classification Speed: 125 sec. per 1,000 documents

Bag-of-words

Hybrid Naïve

Hybrid Smart

Hybrid Smart with Fixed Threshold

Classification Accuracy

Frequent Terms Used

50%  60%  70%  80%  90%  100%

20  30  40  50  60  70  80  90  100
Filtering of Multi-Lingual Terrorist Content

Offline and Online Execution Times for NBC

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Method</th>
<th>Time to Build Graphs (sec)</th>
<th>Time to Build Dictionary (sec)</th>
<th>Time to Construct Vectors (sec)</th>
<th>Time to Build Classification Model (sec)</th>
<th>Total Time Offline (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-series</td>
<td>Hybrid Smart ( (N = 100, CR_{\text{min}} = 1.2) )</td>
<td>223.2</td>
<td>2460.86</td>
<td>4.21</td>
<td>0.12</td>
<td>2688.4</td>
</tr>
<tr>
<td></td>
<td>Hybrid Naïve ( (N = 20, \ t_{\text{min}} = 0.2) )</td>
<td>283.64</td>
<td>1.46</td>
<td>0.5</td>
<td>0.08</td>
<td>285.68</td>
</tr>
<tr>
<td></td>
<td>Hybrid with Fixed Threshold ( (N = 100, \ t_{\text{min}} = 0.1, CR_{\text{min}} = 1.2) )</td>
<td>223.2</td>
<td>62.3</td>
<td>4.19</td>
<td>0.12</td>
<td>289.81</td>
</tr>
<tr>
<td></td>
<td>Bag-of-words ( (N = 100) )</td>
<td>n/a</td>
<td>51.55</td>
<td>286.34</td>
<td>42.62</td>
<td>380.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Method</th>
<th>Average Time to Classify One Document (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-series</td>
<td>Hybrid Smart</td>
<td>( 1.2 \times 10^{-3} )</td>
</tr>
<tr>
<td></td>
<td>Hybrid Naïve</td>
<td>( 6.49 \times 10^{-4} )</td>
</tr>
<tr>
<td></td>
<td>Hybrid with Fixed Threshold</td>
<td>( 5.7 \times 10^{-4} )</td>
</tr>
<tr>
<td></td>
<td>Bag-of-words</td>
<td>0.125</td>
</tr>
</tbody>
</table>
How many subgraphs have more than one node?

 Relative Number of Multi Node Graphs for C4.5, K-series

 Relative Number of Multi Node Graphs for C4.5, U-series

 Relative Number of Multi Node Graphs for NBC, K-series

 Relative Number of Multi Node Graphs for NBC, U-series
Summary of Results

• Different document representations were empirically compared in terms of classification accuracy and execution time

• The hybrid (graph-vector) methods were found to be more accurate in most cases and generally much faster than their vector-space and graph-based counterparts

• The percentage of multi-node subgraphs in the term set was close to 90% in the K-Series and close to 20% in the U-Series
Case Study 1

Categorization of Web Documents in Arabic

(Based on Last et al., 2006)
Document Collection

- **648 Arabic documents**
  - 200 documents downloaded from terrorist web sites
  - 448 belong to non-terrorist categories
- **Terrorist web sites**
  - [http://www.qudsway.com](http://www.qudsway.com) (Palestinian Islamic Jihad)
- **Normal (non-terrorist) web sites**
  - [www.aljazeera.net/News](http://www.aljazeera.net/News)
  - [http://arabic.cnn.com](http://arabic.cnn.com)
Preprocessing of Documents in Arabic

• Normalizing orthographic variations
  – E.g., convert the initial Alif Hamza ء to plain Alif ﬁ

• Normalize the feminine ending, the Ta-Marbuta ﯥ, to Ha ﯤ

• Removal of vowel marks

• Removal of certain letters (such as: Waw و, Kaf ك, Ba ب, and Fa ﯻ) appearing before the Arabic article THE (Alif + Lam ﯤ)

• Removal of pre-defined stop words in Arabic

• Final vocabulary size: 47,836 words
Accuracy Results

Results for Naïve Approach with C4.5 Classifier

Results for Smart Approach with C4.5 Classifier
Filtering of Multi-Lingual Terrorist Content

Resulting Decision Tree

الصهيوني
The Zionist (Adj., Sing. M.)

Yes

No

Terror

الشهيد
The Martyr

Yes

No

Terror

الصهيوني
The Zionist (Adj.,
Sing. F. or Pl.)

Yes

No

Terror

نداء
Call

القدس
Al-Quds

Yes

No

العدو
The Enemy

Yes

No

Terror

Non-Terror
Does the word الحركة الصهيونية ("Zionist") indicate a terrorist document?

- The word "Zionist" occurred only in six normal documents out of 448.
- It never occurred more than once in the same normal document.
- On normal documents, the word was used in the following expressions:
  - The Zionist Movement - الحركة الصهيونية
  - The Zionist enemies - العدوان الصهيوني
  - The Zionist plot - المؤامرة الصهيونية
  - The Zionist extremists - غلاة الصهيونية
  - The First Zionist Congress - المؤتمر الصهيوني الأول
  - The extremist Zionist groups - الجماعات الصهيونية المتطرفة
Case Study 2

Categorization of Terrorist Web Documents in English
Document Collection

• 1,004 English documents
  – 913 documents downloaded from a Hezbollah website (http://www.moqawama.org/english/)
  – 91 documents downloaded from a Hamas website (www.palestine-info.co.uk/am/publish/)

• Goal
  – Identify the source of web documents (Hamas vs. Hezbollah)

• Document Representation
  – The Hybrid Smart approach

• Classifier
  – C4.5 Decision Tree
Accuracy Results

Maximum Graph Size: 100 Nodes

Subgraph Frequency Threshold

Classification Accuracy (%)

Tree Size

Accuracy (%) vs. Subgraph Frequency Threshold

Tree Size (Nodes) vs. Subgraph Frequency Threshold

Accuracy:
99.10 99.10 99.10 99.10 99.10

Tree Size:
11 11 11 11 11 9

0.30 0.35 0.40 0.45 0.50 0.55 0.60

Mark Last (BGU)

Filtering of Multi-Lingual Terrorist Content

December 19, 2006
Resulting Decision Tree
Subgraph Frequency Threshold: 0.55
Conclusions

• Automated filtering of multi-lingual terrorist content is a feasible task
  – Graph representations contribute to categorization accuracy
  – Hybrid (graph and vector) methods improve the processing speed
  – Decision trees provide an interpretable structure that can be tested by a human expert
Future Work

• Some open challenges
  – Developing graph representations of web documents for more languages
  – Finding optimal parameters for subgraph extraction
  – Multi-label categorization of terrorist documents
  – Improving classification accuracy using ontologies of the terrorist domain
  – Identification of groups and topics
References (1)

References (2)

- M. Kuramochi and G. Karypis. An Efficient Algorithm for Discovering Frequent Subgraphs. *IEEE Transactions on Knowledge and Data Engineering* 16, 9 (Sep. 2004).
References (3)