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

http://www.ise.bgu.ac.il/faculty/mlast/presentations/icdm2006_fmtc.pdf

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

Sabiroon - Hamas

Language: English

Palestine Info – Hamas

Language: French

Palestine Info – Hamas

Language: Russian

Qudsway – Palestinian Islamic Jihad

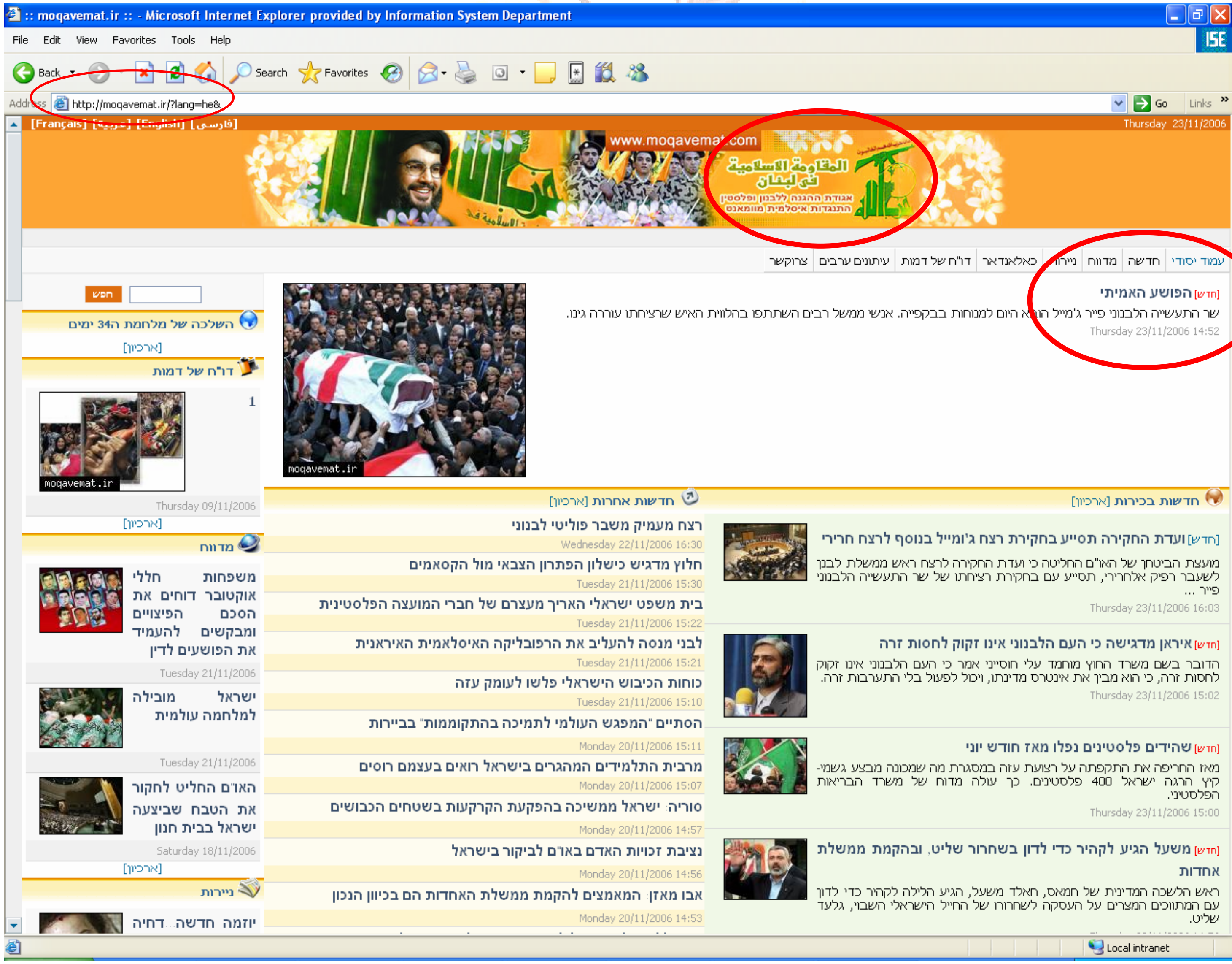
Language: Arabic

Army of Ansar Al-Sunna (Iraq)

Language: Arabic

Hezbollah (Lebanon)

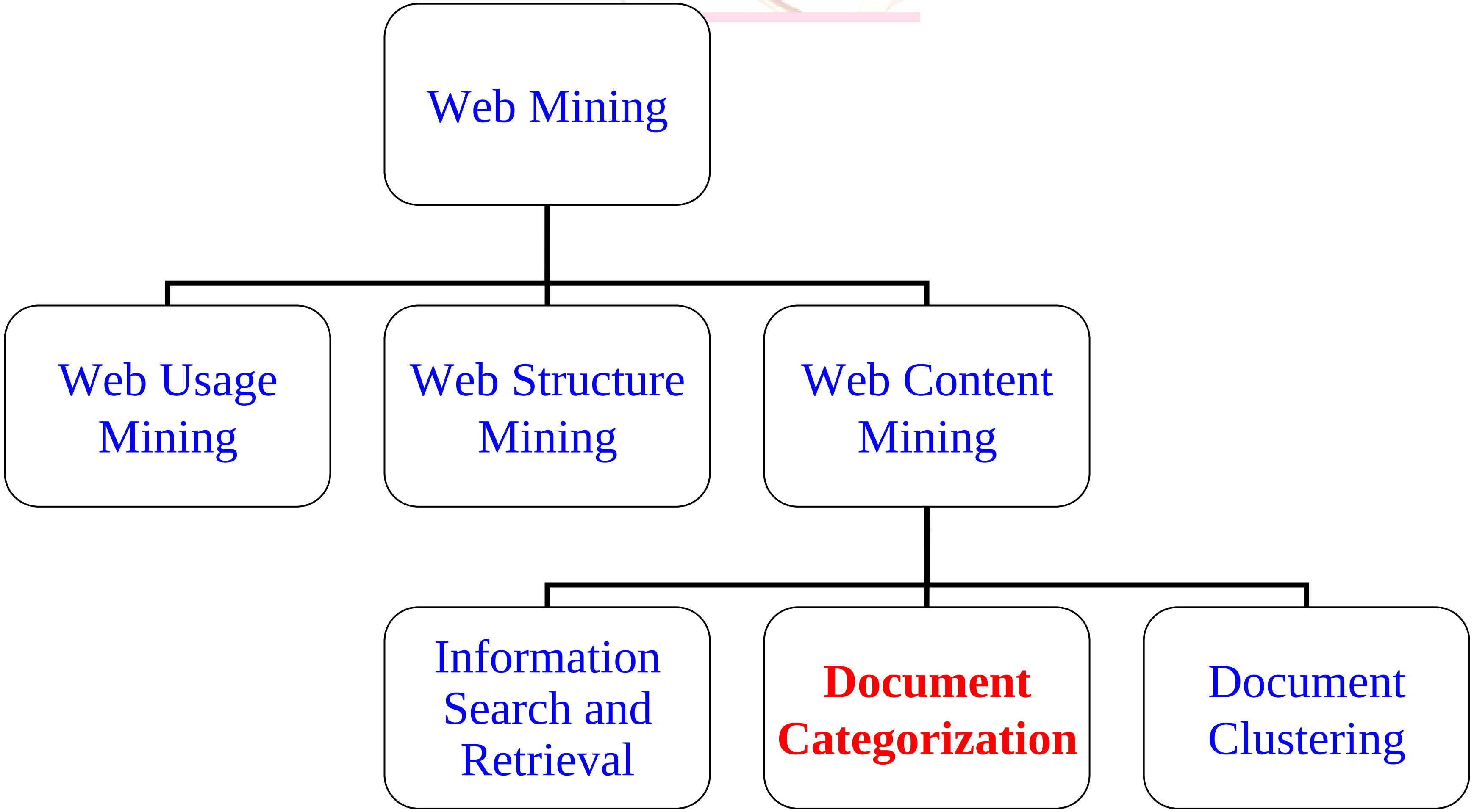
Language: Hebrew



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)

Web Mining Tasks



Text Categorization (TC)
Basic Definition

- TC – task of assigning a Boolean {T, F} value to each pair $\langle d_j, c_i \rangle \in D \times C$

  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 $d_j = (w_{1j},\dots,w_{|T|j})$ where $T = (t_1,\dots,t_{|T|})$ 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

$$w_{ij} = \frac{TF_{ij}}{\max\limits_j TF_{ij}}$$

where TF_{ij} = raw frequency of term t_j in document d_i

- TFIDF (term frequency × inverse document frequency)

$$w_{ij} = TF \times IDF = TF_{ij} \times \log \frac{N}{n}$$

where

N = number of documents in collection (corpus)

n = number of documents where term t_j occurs at least once

The “Bag of Words” Approach
A Practical Example

Text 1

From palestine-info.co.uk
Dec 10, 2005
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.

↓

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

Text 2

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

Terrorist

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!

Non-Terrorist

Bag of Words 2

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

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

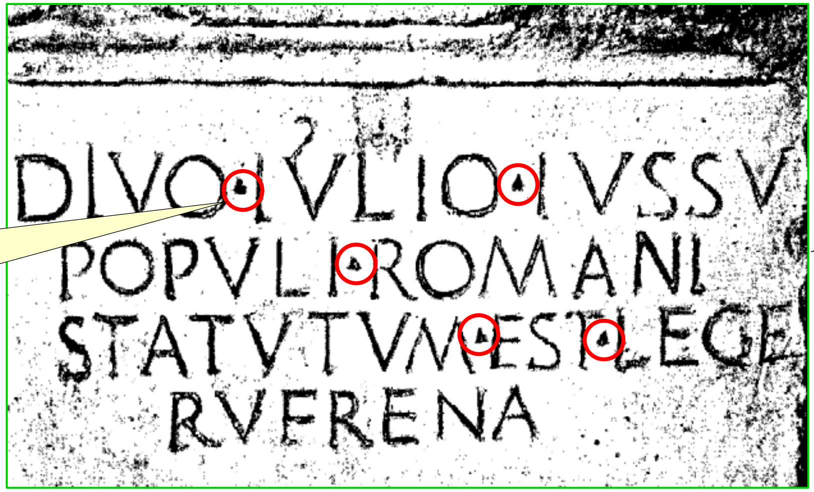
DIVIDE ET IMPERA

("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:

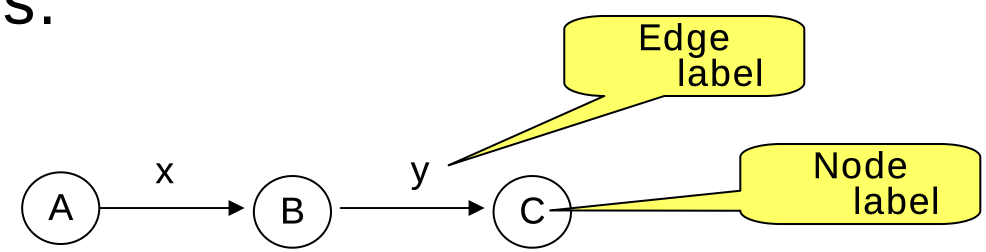
The Graph-Based Model

(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, α is a function labeling the nodes and β is a function labeling the edges.



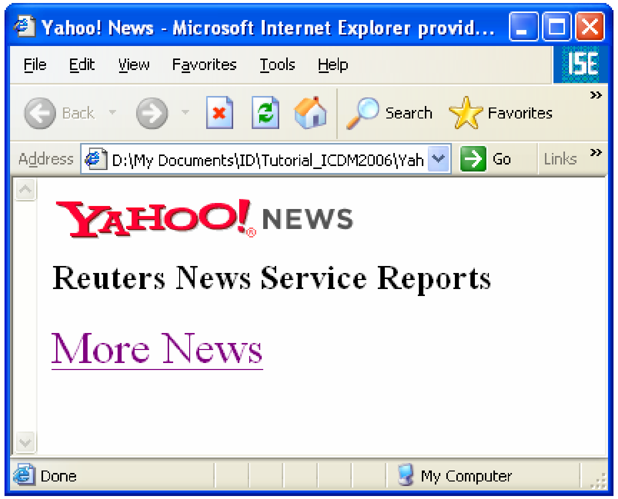
- Node and edge IDs are omitted for brevity
- **Graph size**: $|G| = |V| + |E|$

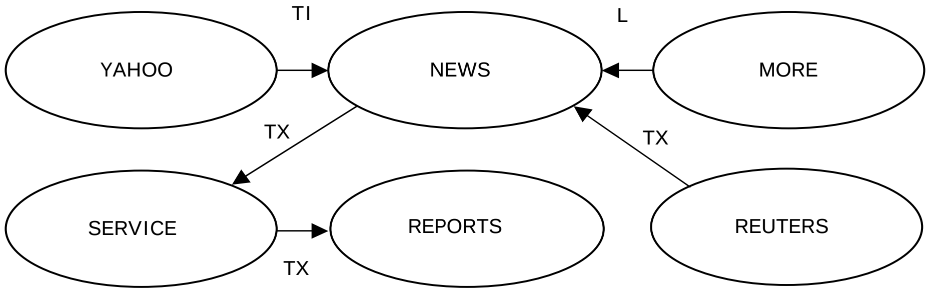
The Graph-Based Model of Web Documents

- 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

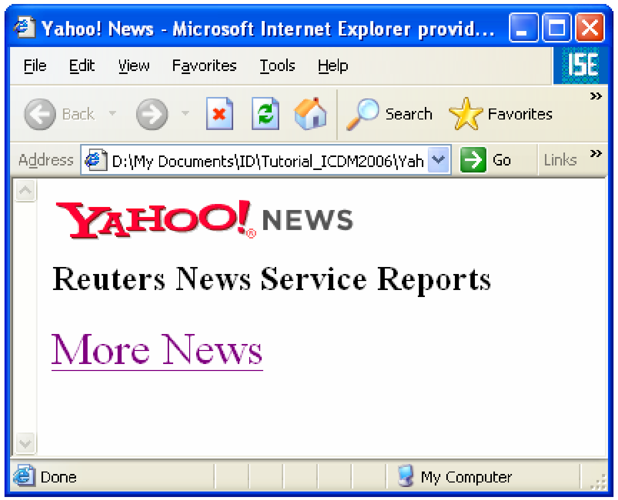
- Edges are labeled according to the document section where the words are followed by each other
  - *Title (TI)* contains the text related to the document's title and any provided keywords (meta-data);
  - *Link (L)* is the "anchor text" that appears in clickable hyper-links on the document;
  - *Text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

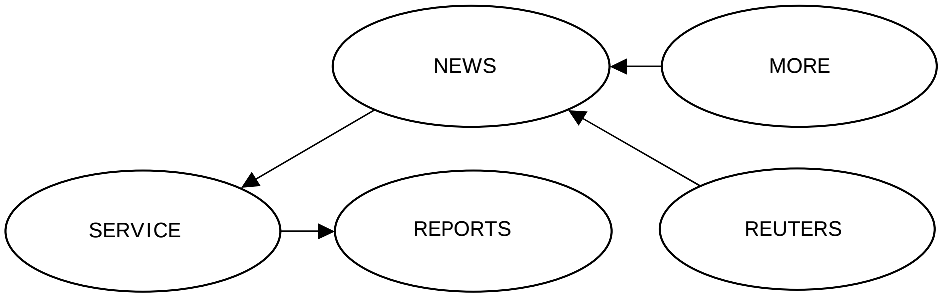




The *Simple* Representation

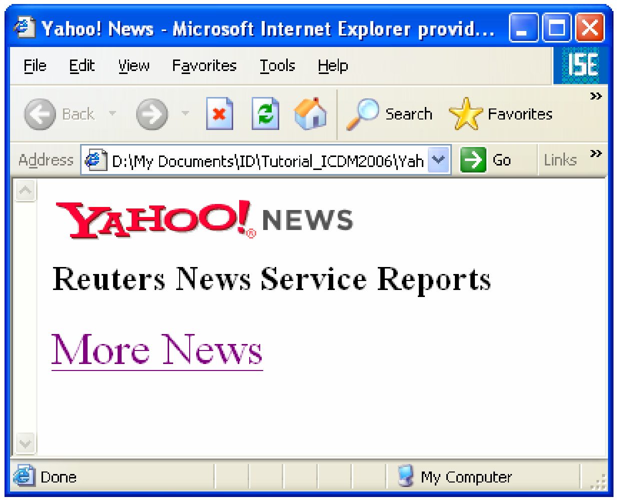
- The graph is based only the visible text on the page (title and meta-data are ignored)
- Edges are not labeled



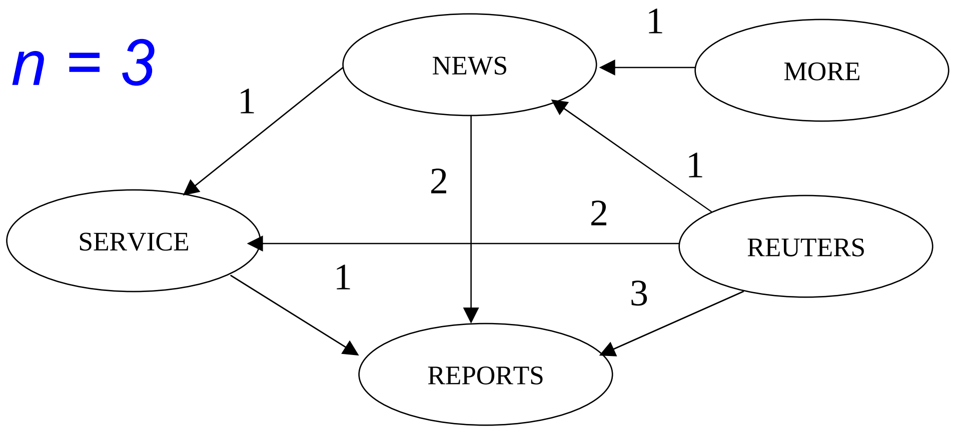


The *n-distance* Representation

- 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.

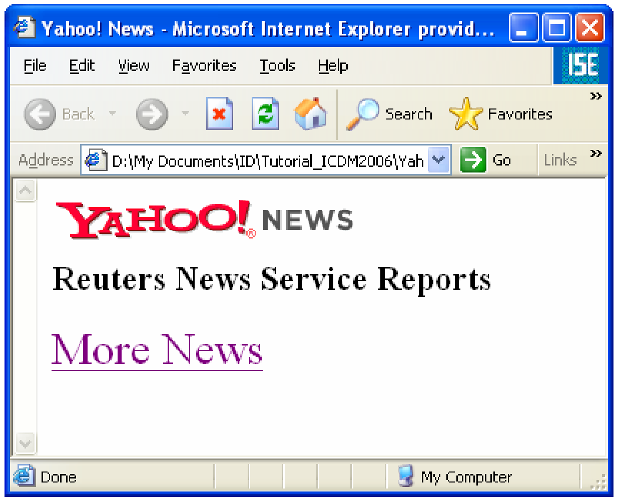


n = 3

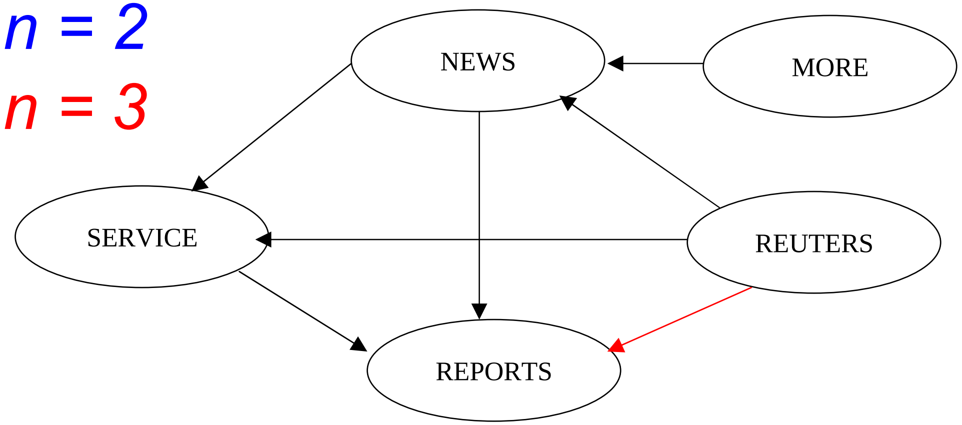


The *n-simple* Representation

- Based on the visible text only
- We look up to *n* terms ahead and connect the succeeding terms with an <u>unlabeled</u> edge
- *n* is a user-provided parameter.

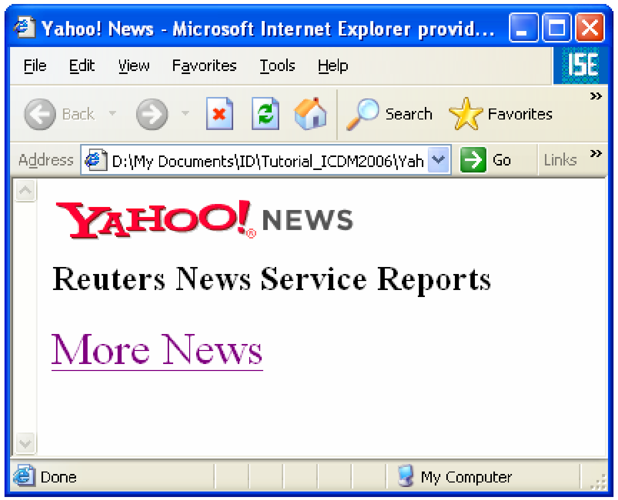


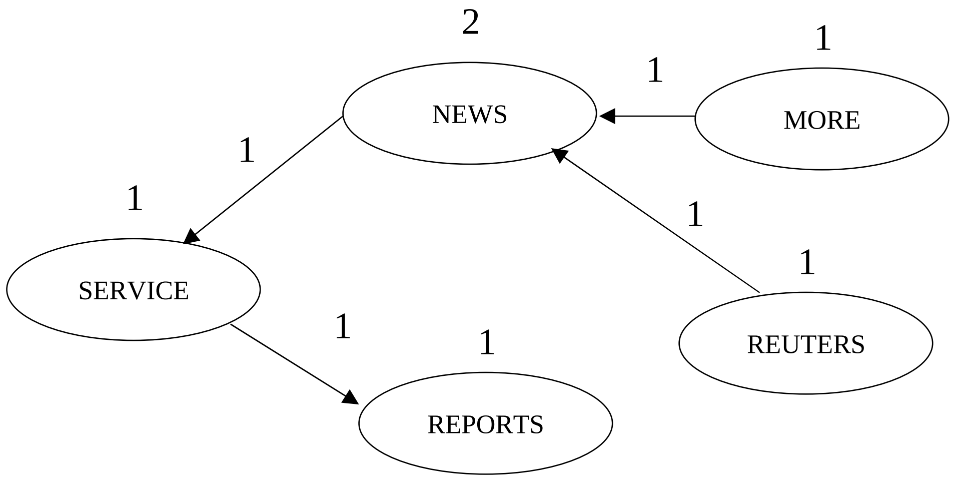
n = 2
n = 3



The *Absolute Frequency* Representation

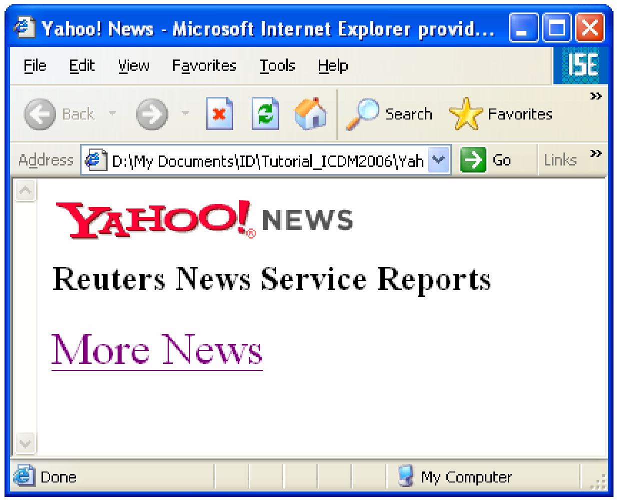
- No section-related information
- Each node and edge is labeled with an absolute frequency measure

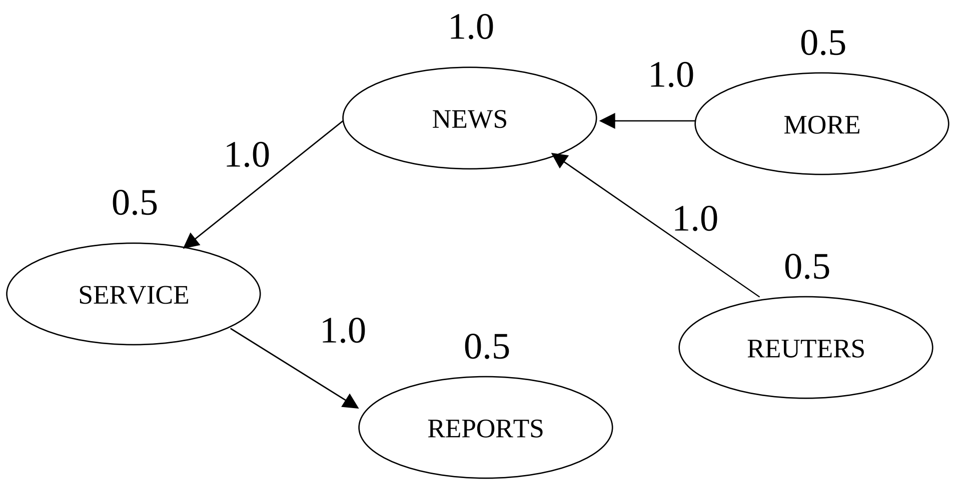




The *Relative Frequency* Representation

- 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





Graph Based Document Representation – Detailed Example

Source: www.cnn.com, May 24, 2005

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

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//E
<!-- saved from url=(0023)http://edition.cnn.com/ -->
<HTML lang=en><HEAD><TITLE>CNN.com International</TITLE>
<META http-equiv=content-type content="text/html; charset=iso-8859-1">
<META http-equiv=refresh content=1800><LINK href="/" rel=Start><LINK


<DIV class=cnnSectionT1
style="PADDING-RIGHT: 6px; PADDING-LEFT: 6px; PADDING-BOTTOM: 3px; PADDING-TOP: 3px">
<H2><A style="COLOR: #000"
href="http://edition.cnn.com/2005/WORLD/meast/05/23/iraq.main/index.html">Iraq
bomb: Four dead, 110 wounded</A></H2>
<P>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.</P>
<P><A class=cnnt1link
href="http://edition.cnn.com/2005/WORLD/meast/05/23/iraq.main/index.html">FULL
STORY</A></P>
```

title

link

text

Graph Based Document Representation - Preprocessing

TITLE

CNN.com International

Stop word removal

Text

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.

Stemming

Links

Iraq bomb: Four dead, 110 wounded.

FULL STORY.

Graph Based Document Representation - Preprocessing

TITLE

CNN.com International

Text

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 Iraqis Prime Minister Ibrahim al-Jaafari and his driver were killing in a driver shooting.

Links

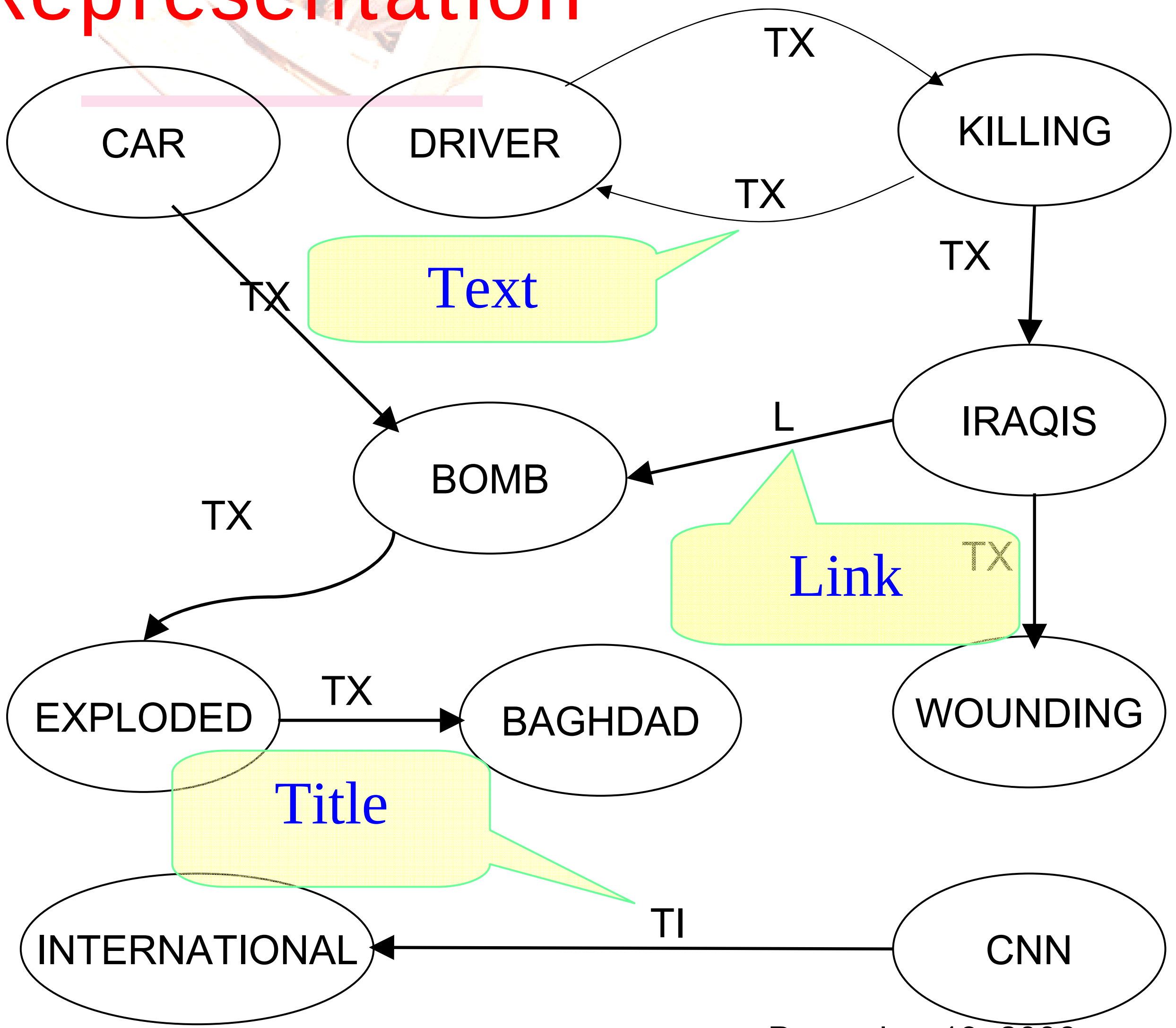
Iraqis bomb: Four dead, 110 wounding.

FULL STORY.

Standard Graph Based Document Representation

Ten most frequent terms are used

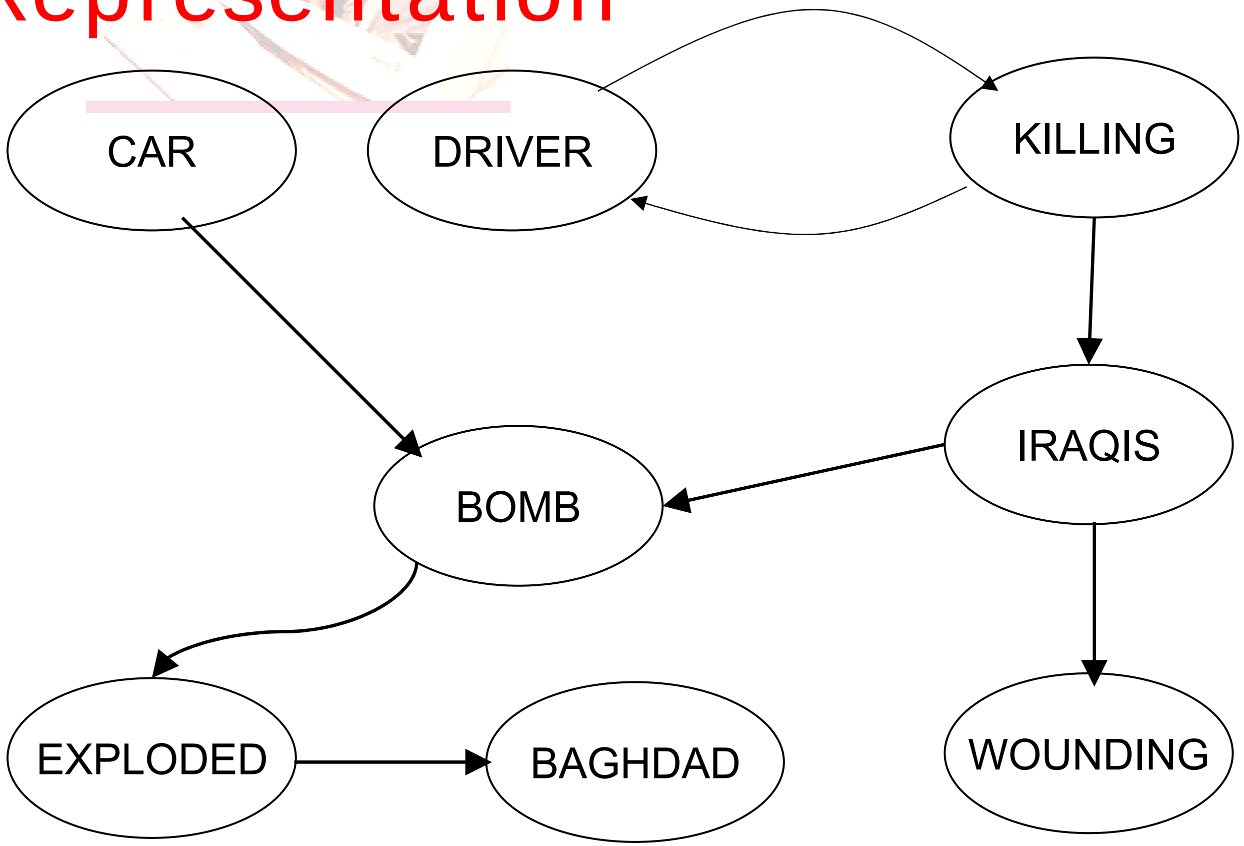
Word	Frequency
Iraqis	3
Killing	2
Bomb	2
Wounding	2
Driver	2
Exploded	1
Baghdad	1
International	1
CNN	1
Car	1



Simple Graph Based Document Representation

Ten most frequent terms are used

Word	Frequency
Iraqis	3
Killing	2
Bomb	2
Wounding	2
Driver	2
Exploded	1
Baghdad	1
International	1
CNN	1
Car	1



"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

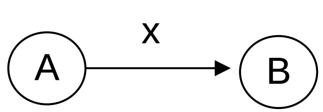
- Required properties
  - (1) *boundary condition*: $d(G_1,G_2) \geq 0$
  - (2) *identical graphs have zero distance*: $d(G_1,G_2)=0 \rightarrow G_1 \cong G_2$
  - (3) *symmetry*: $d(G_1,G_2)=d(G_2,G_1)$
  - (4) *triangle inequality*: $d(G_1,G_3) \leq d(G_1,G_2)+d(G_2,G_3)$

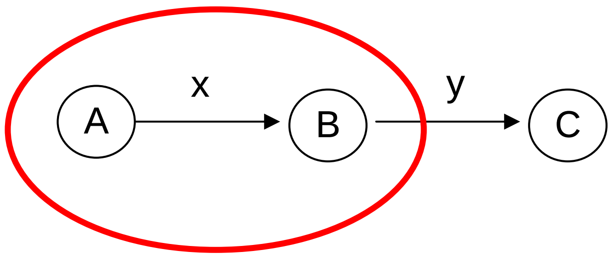
Relevant Definitions

(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

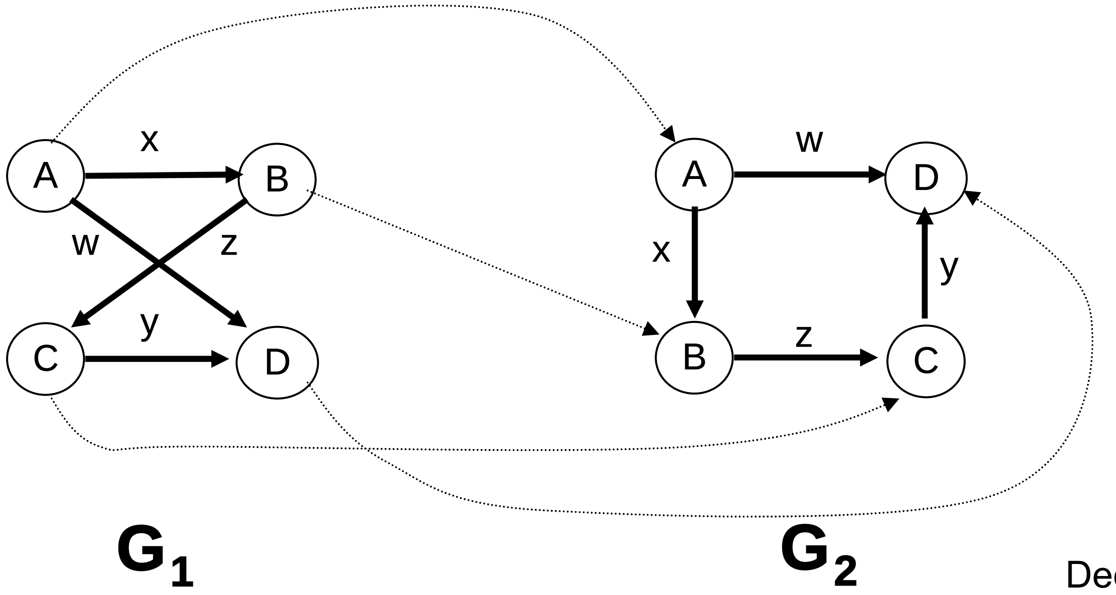




G_1 **G_2**

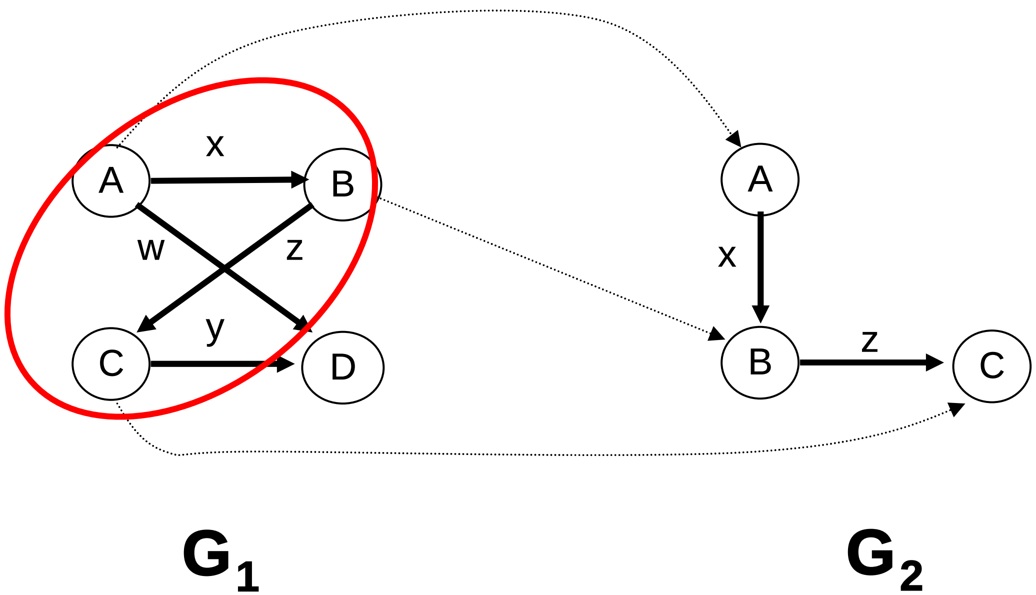
More Graph-Theoretic Definitions

- 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 $\beta_1(x, y) = \beta_2(f(x), f(y))$ $\forall (x, y) \in V_1 \times V_1$.



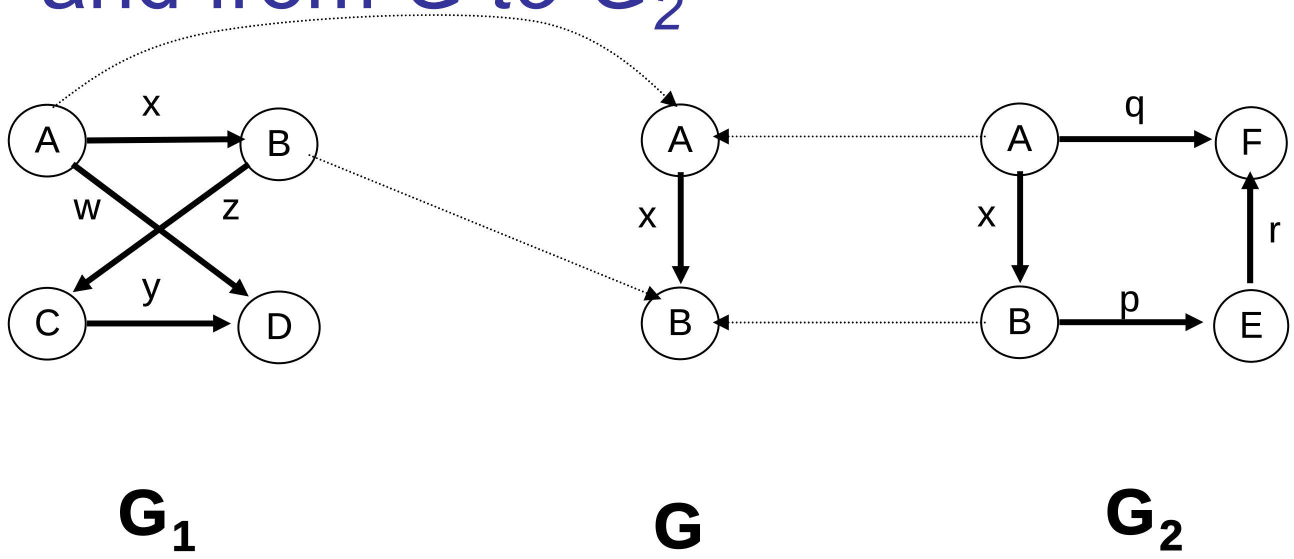
More Graph-Theoretic Definitions

- **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.



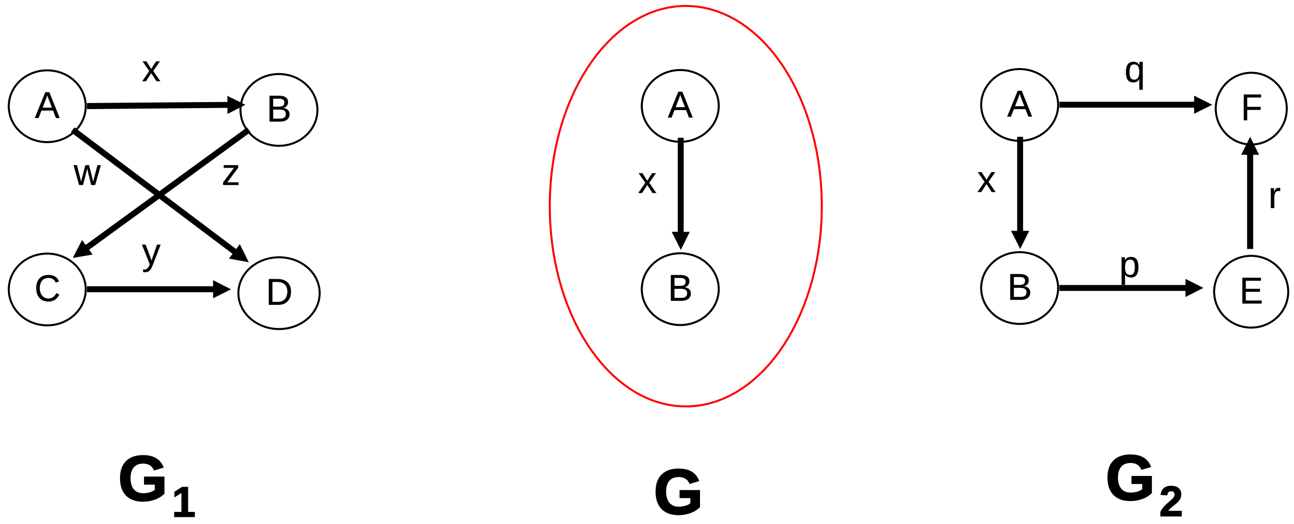
More Graph-Theoretic Definitions

- 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



More Graph-Theoretic Definitions (cont.)

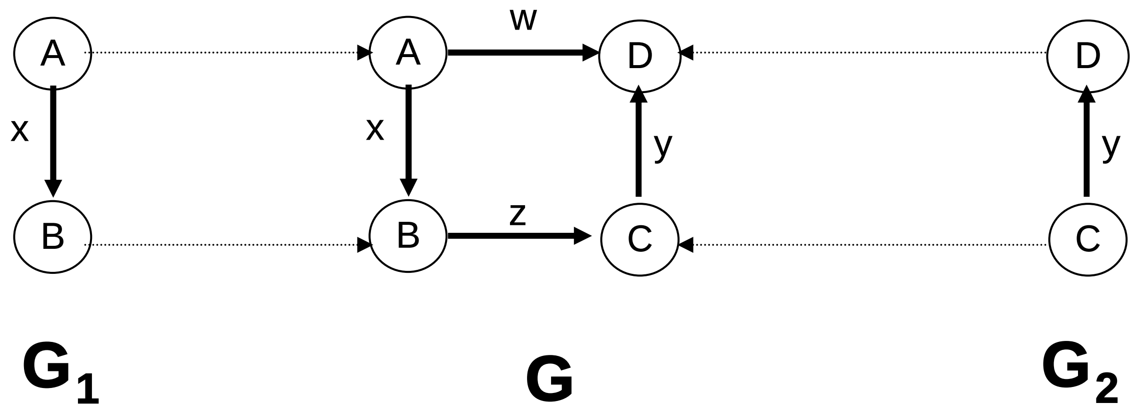
- The graph G is a **maximum common subgraph (mcs)** if G is a common subgraph of G_1 and G_2 and there exist no other common subgraph G' of G_1 and G_2 such that $|G'| > |G|$



$$|G| = |V| + |E| = 2+1 = 3$$

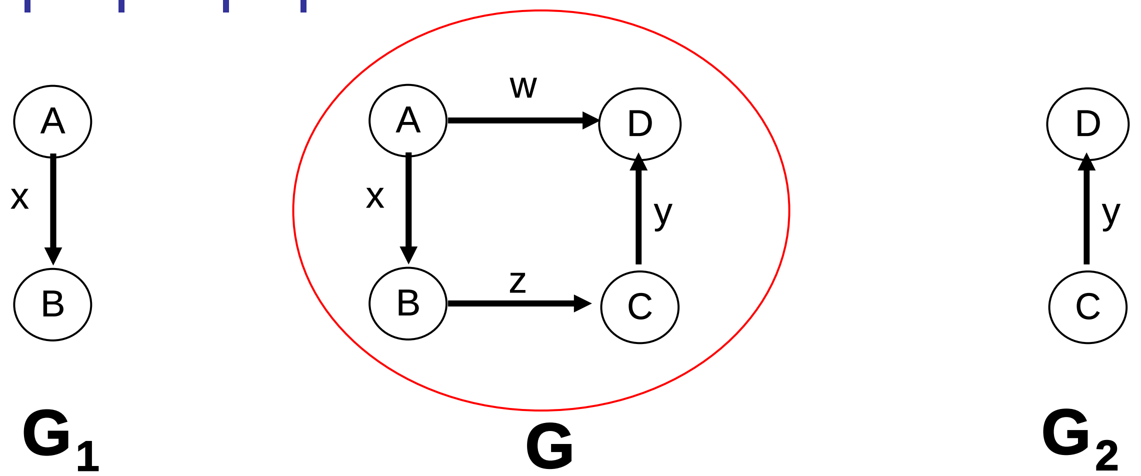
More Graph-Theoretic Definitions (cont.)

- 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 $|G'| < |G|$



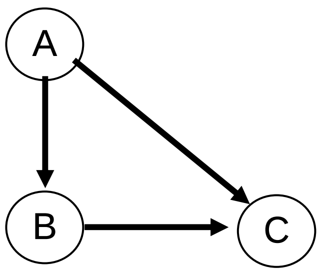
$$|G| = |V| + |E| = 4 + 2 = 6$$

Distance between two Graphs

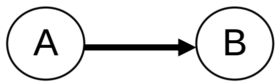
- MMCSN Measure (Schenker et al., 2005):

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

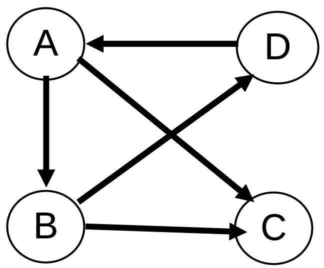
- *mcs(G_1, G_2)* - maximum common subgraph
- *MCS(G_1, G_2)* - minimum common supergraph



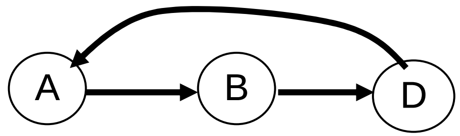
G_1



mcs (G_1,G_2)



MCS (G_1,G_2)



G_2

$$d_{MMCSN}(G_1,G_2)=1-\frac{2+1}{4+5}=0.667$$

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):

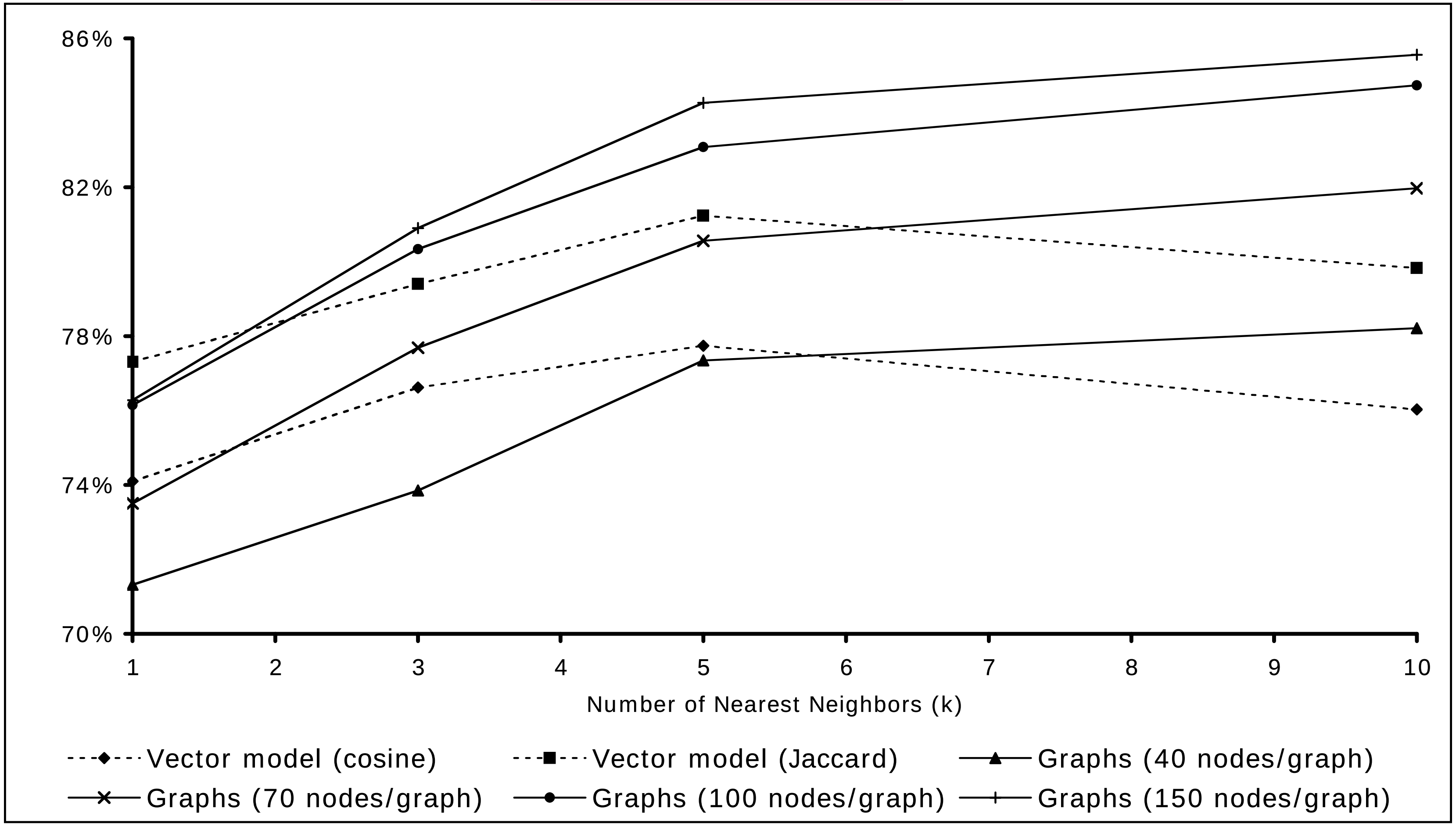
$$d_{MMCS}(G_1,G_2) = |MCS(G_1,G_2)| - |mcs(G_1,G_2)|$$

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: ftp://ftp.cs.umn.edu/dept/users/boley/PDDPdata/
  - 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*

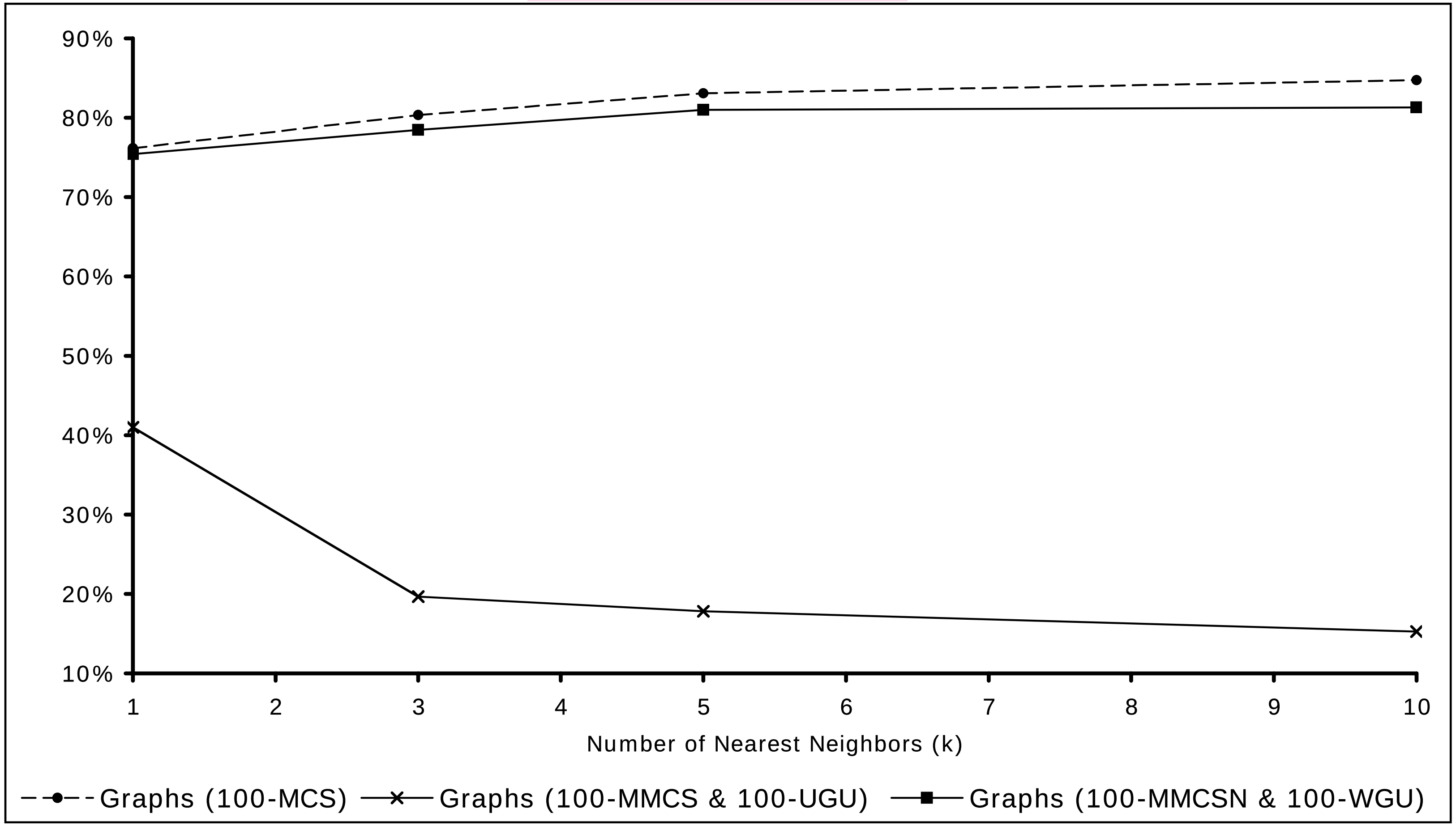
k-Nearest Neighbors with Graphs

Accuracy vs. Graph Size



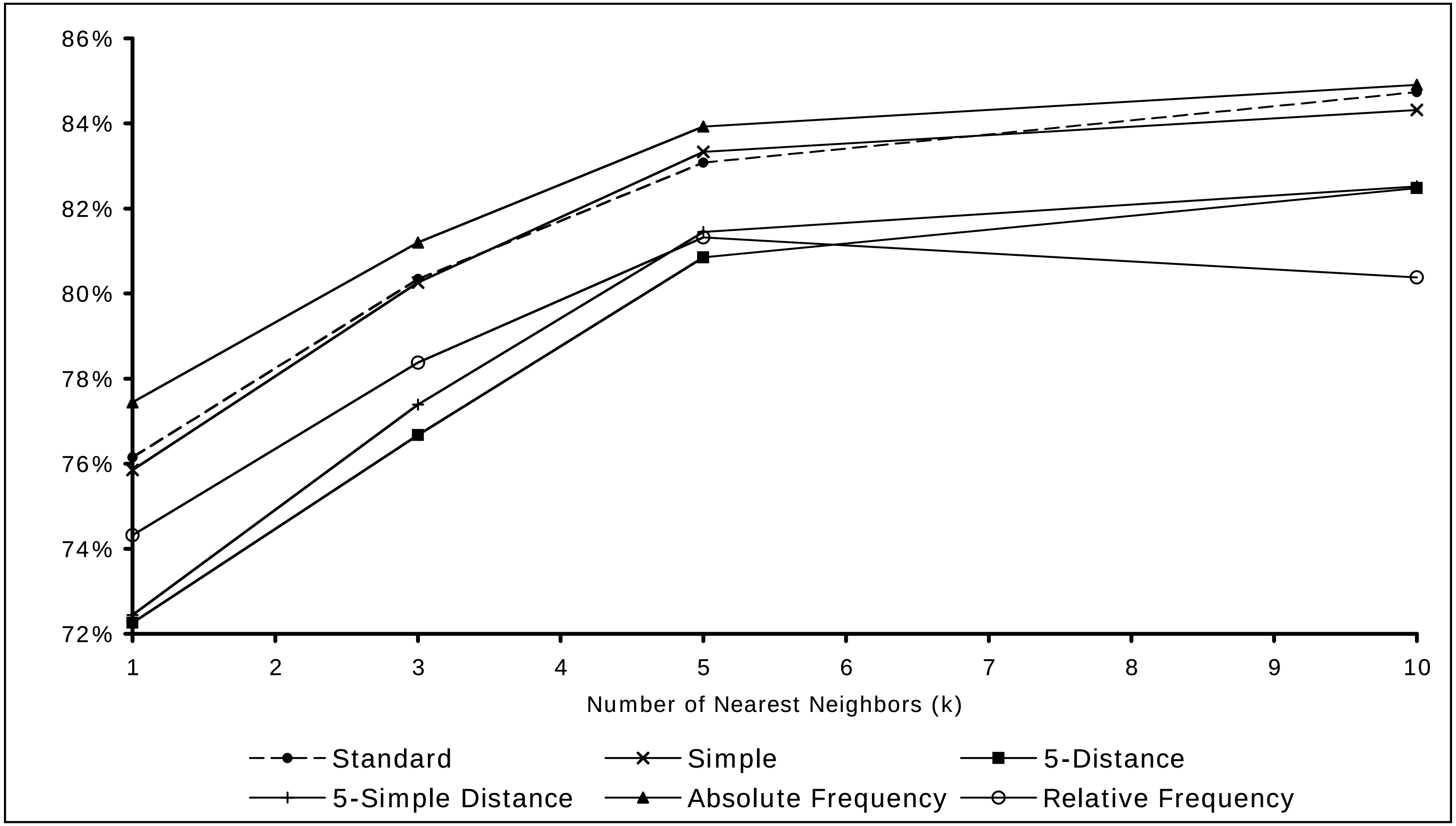
k-Nearest Neighbors with Graphs

Accuracy vs. Distance Measure



k-Nearest Neighbors with Graphs

Accuracy vs. Graph Representation



k-Nearest Neighbors with Graphs

Average Time to Classify One Document

Method	Average time to classify one document
Vector (cosine)	7.8 seconds
Vector (Jaccard)	7.79 seconds
Graphs, 40 nodes/graph	8.71 seconds
Graphs, 70 nodes/graph	16.31 seconds
Graphs, 100 nodes/graph	24.62 seconds

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)

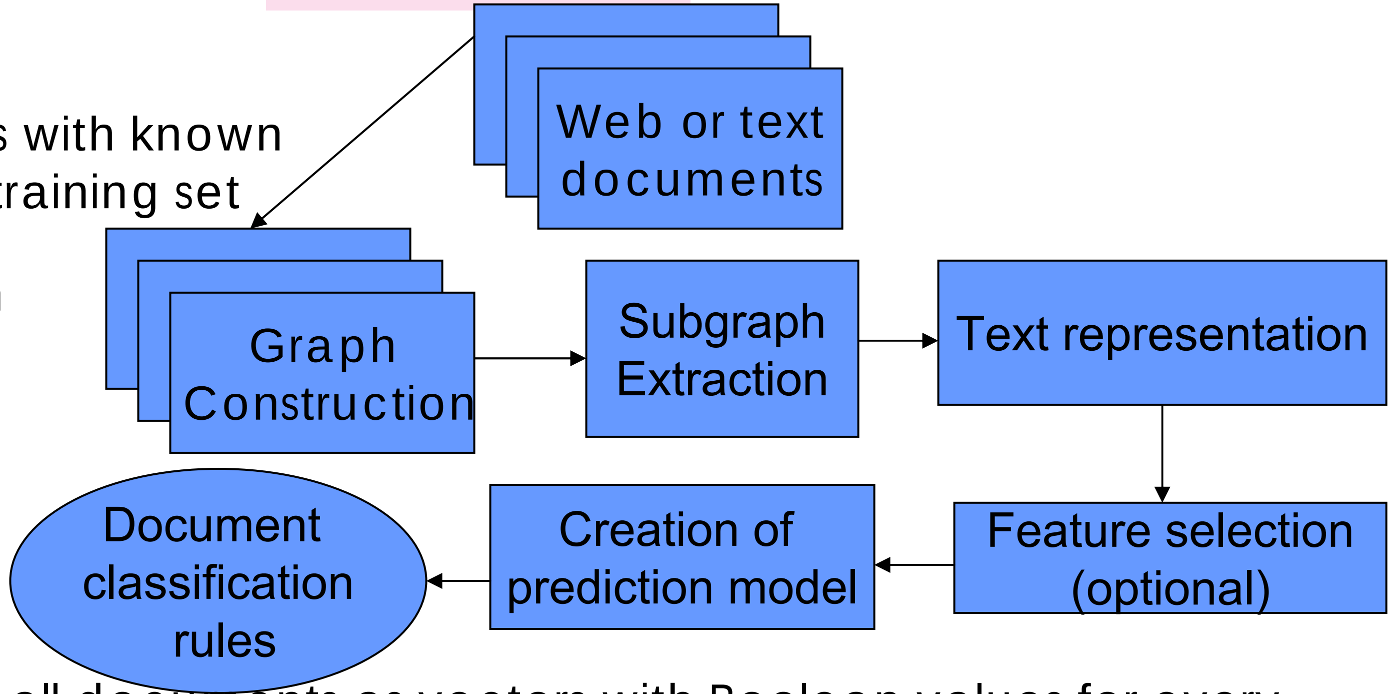
- Basic Idea
  - Represent a document as a vector of sub-graphs
  - Categorize documents with a *model-based classifier* (e.g., a decision tree), which is much faster than a "lazy" method
- Naïve Approach
  - Select sub-graphs that are most frequent in each category
- Smart Approach
  - Select sub-graphs 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***



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_{min} – Threshold (minimum subgraph frequency)
- Output:
  - A set of classification-relevant subgraphs
- Process:
  - For each category, find frequent subgraphs $SCF > t_{min}$
  - *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*** > $\boldsymbol{CR}_{\text{min}}$
  - 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):

$$SCF\left(g'_k(c_i)\right) = \frac{g'_k f(c_i)}{N(c_i)}$$

$SCF\left(g'_k(c_i)\right)$ - frequency of sub-graph g'_k in category c_i

$N(c_i)$ - Number of documents in category c_i

$g'_k f(c_i)$ - Number of documents containing g'_k in category c_i

The Smart Subgraph Extraction (cont.)

- Inverse Subgraph Frequency:

$$ISF\left(g'_k(c_i)\right)=\begin{cases}\log_2\left(\dfrac{\sum N(\mathrm{c_j})}{\sum g'_k f\,(\mathrm{c_j})}\right) & \text{if } \sum g'_k f\,(\mathrm{c_j})>0\\ \log_2\left(2\times\sum N(\mathrm{c_j})\right) & \text{if } \sum g'_k f\,(\mathrm{c_j})=0\end{cases}\qquad \{\forall c_j\in C;\; j\neq i\}$$

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

$N(c_j)$ - Number of documents in category c_j

$g'_k f\,(c_j)$ - Number of documents containing g'_k in category c_j

The Smart Subgraph Extraction (cont.)

- Subgraph Classification Rate:

$$CR\left(g'_k(c_i)\right) = SCF\left(g'_k(c_i)\right) \times ISF\left(g'_k(c_i)\right)$$

- *$SCF(g'_k(c_i))$* - 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
  - $t_{\min}$ – Threshold (minimum subgraph frequency)
  - CR_{min} - Minimum Classification Rate
- Output
  - Set of classification-relevant subgraphs
- Process:
  - For each class find subgraphs $\boldsymbol{SCF > t_{\min}}$ and $\boldsymbol{CR > CR_{\min}}$
  - 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

Notation	Description
G	Set of document graphs
t_{min}	Subgraph frequency threshold
K	Number of edges in the graph
G	Single graph
sg	Single subgraph
sg^k	Subgraph with k edges
F^k	Set of frequent subgraphs with k edges
E^k	Set of extension subgraphs with k edges
C^k	Set of candidate subgraphs with k edges

Frequent Subgraphs Extraction: The Naïve Algorithm

(based on the FSG algorithm by Kuramochi and Karypis, 2004)

$F^0 \leftarrow$ Detect all frequent single node subgraphs (nodes) in G

$k \leftarrow 1$

While $F^{k-1} \neq \emptyset$ **Do**

  For Each subgraph $sg^{k-1} \in F^{k-1}$ **Do**

    For Each graph $g \in G$ **Do**

      If sg^{k-1} is subgraph of g **Then**

      $E^k \leftarrow$ Detect all possible k edge extensions of sg^{k-1} in g

  For Each subgraph $sg^k \in E^k$ **Do**

    If sg^k already a member of C^k **Then**

      $\{sg^k \in C^k\}.Count{+}{+}$

    Else

      $sg^k.Count \leftarrow 1$

      $C^k \leftarrow sg^k$

  $F^k \leftarrow \{sg^k \text{ in } C^k \mid sg^k.Count > t_{min} * |G|\}$

  $k{+}{+}$

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 $V_1 \subseteq V_2$ - $O(|V_1|*|V_2|)$
2. Determine that $E_1 \subseteq E_2$ – $O(|V_1|^2)$

Total complexity:

$O(|V_1|*|V_2| + |V_1|^2) \leq O(|V_2|^2)$

Graph isomorphism

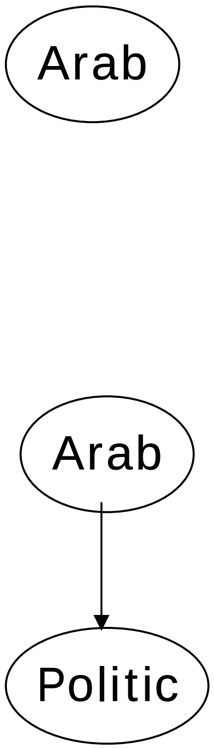
Isomorphism between graphs $G_1=(V_1,E_1,\alpha_1,\beta_1)$ and $G_2=(V_2,E_2,\alpha_2,\beta_2)$ can be found by two simple actions:

1. Determine $G_1 \subseteq G_2$ - $O(|V^2|)$
2. Determine $G_2 \subseteq G_1$ - $O(|V^2|)$

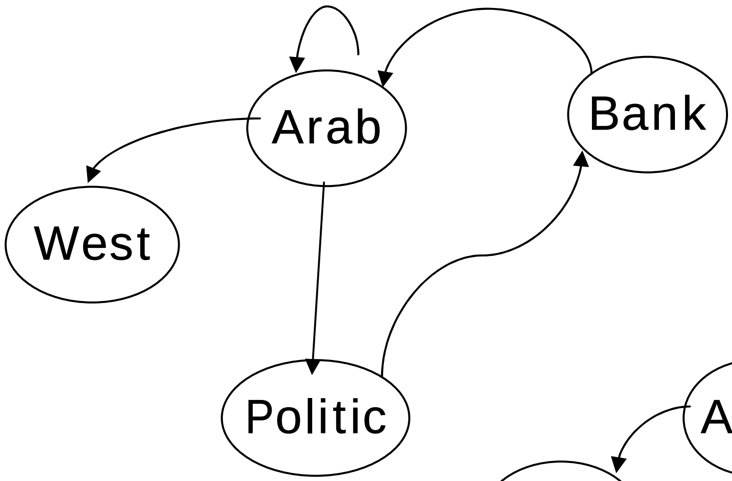
Total complexity: $O(|V^2|)$

Frequent Subgraph Extraction Example

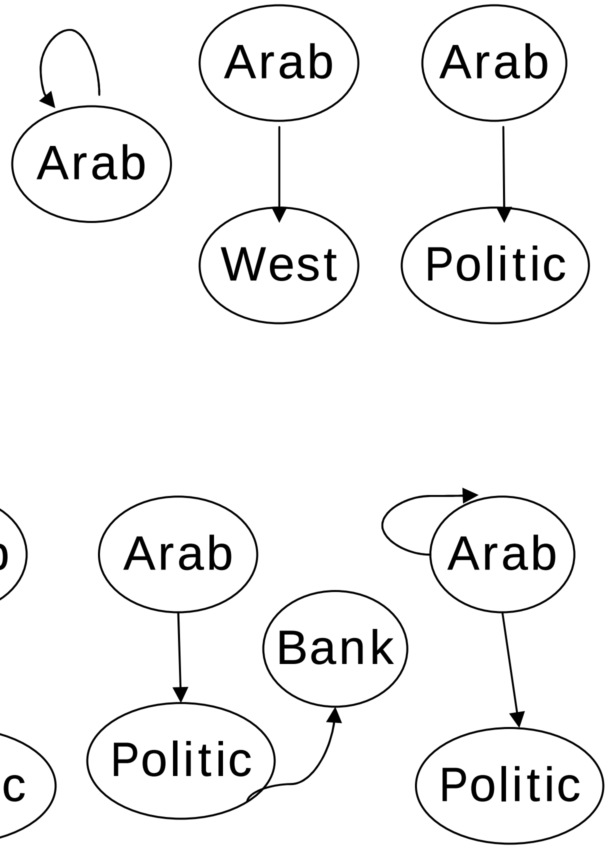
Subgraphs



Document Graph



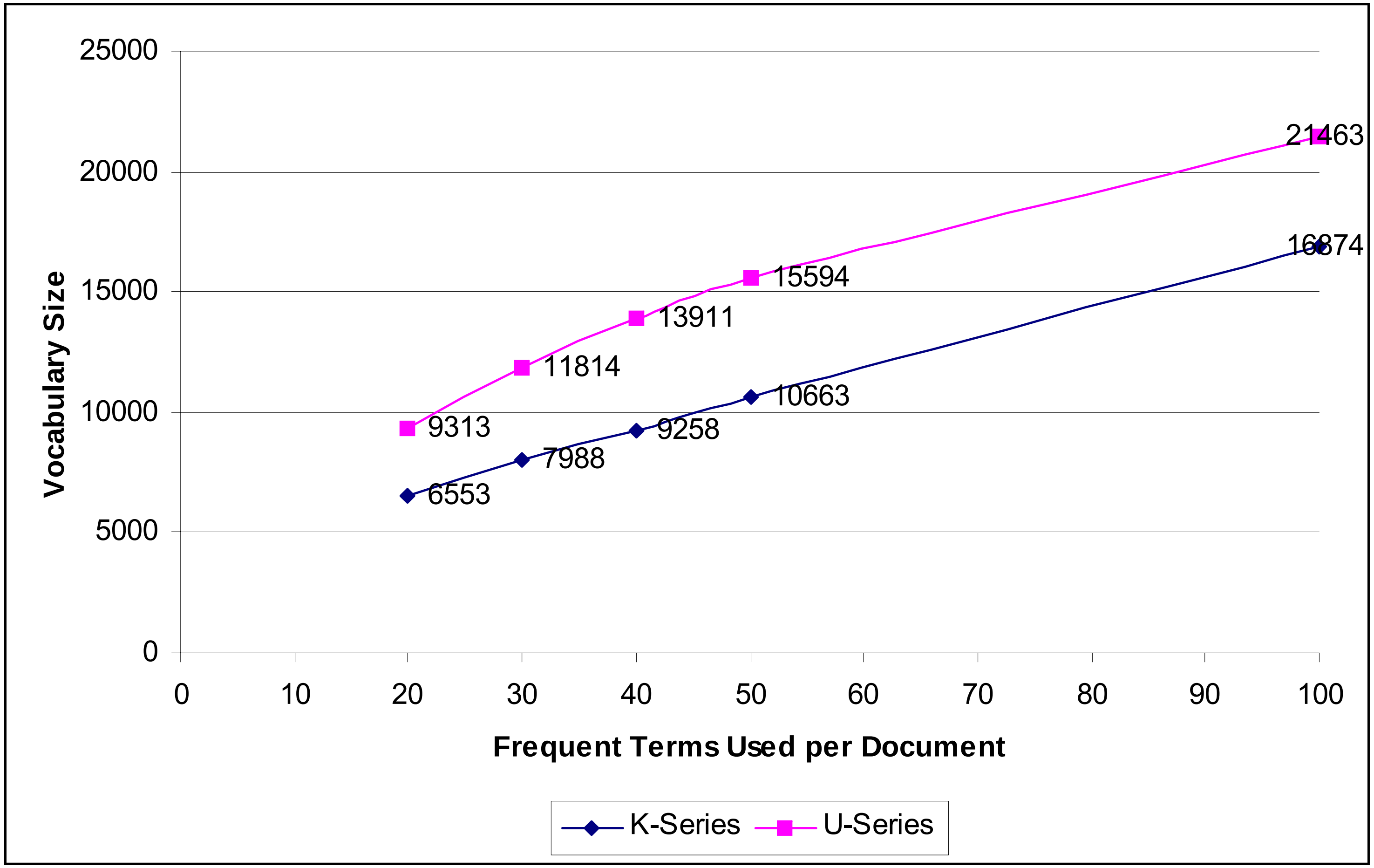
Extensions



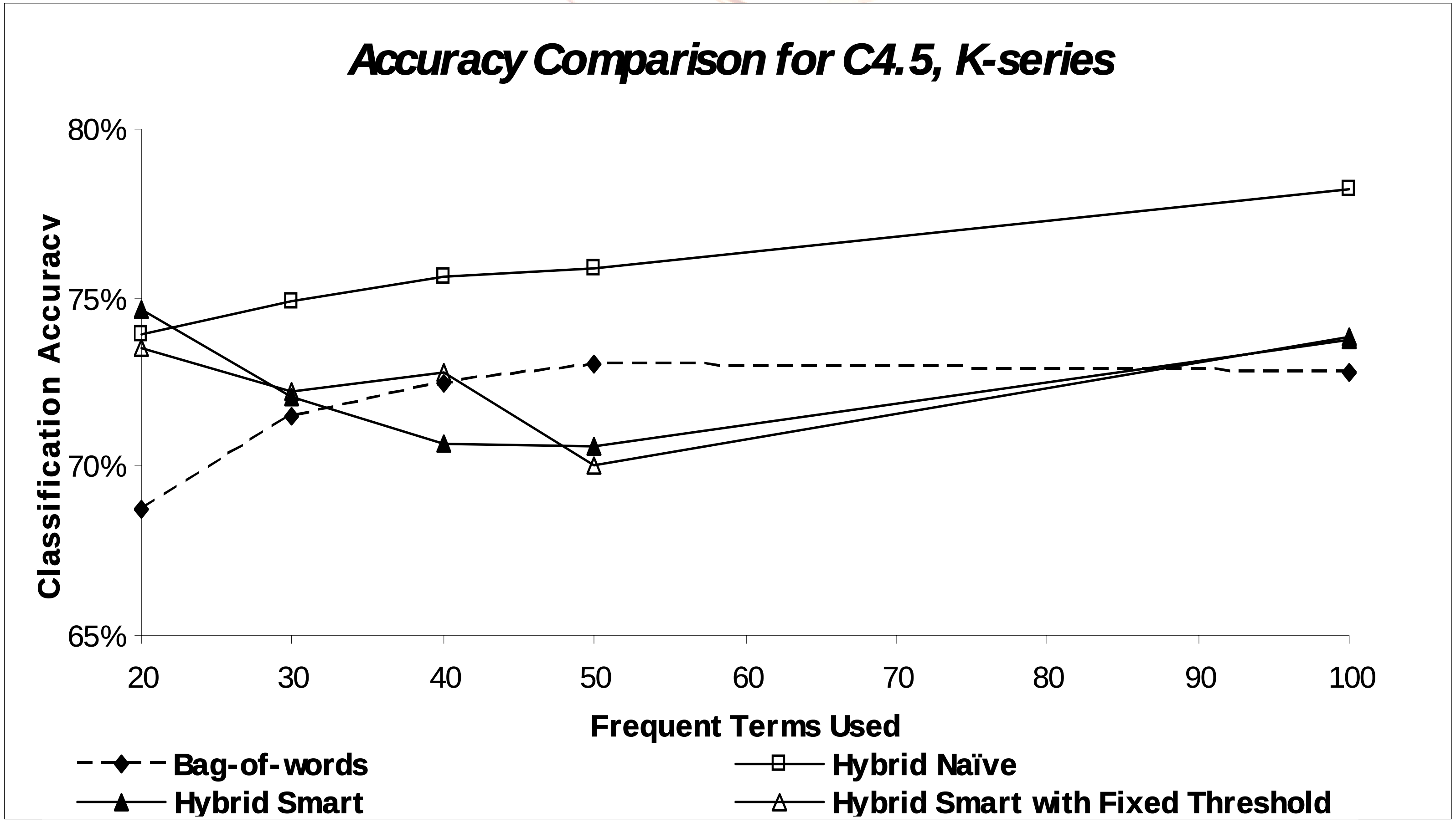
Comparative Evaluation

- 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

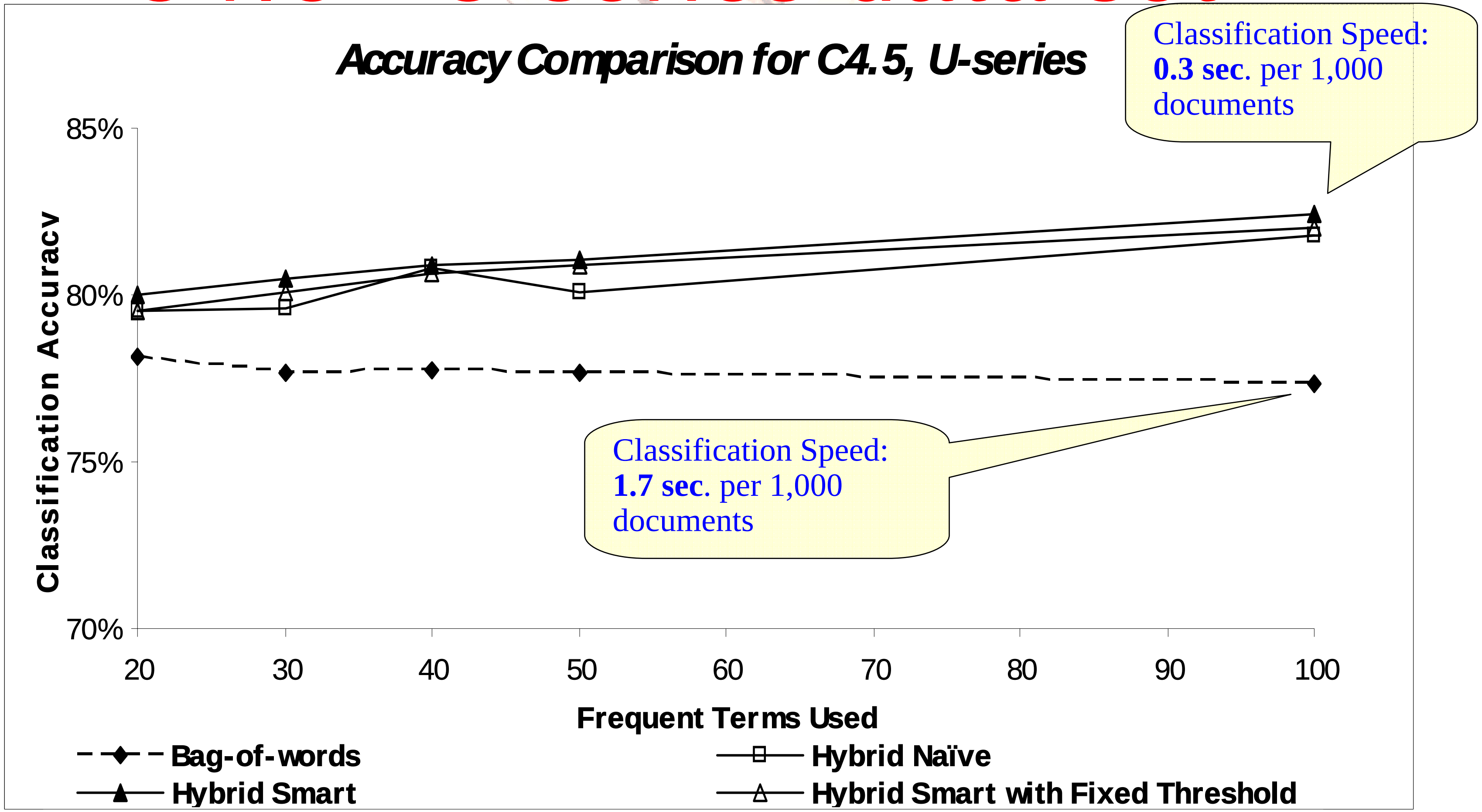
Vocabulary Size as a Function of Frequent Terms Used



Classification Results with C4.5– K series data set



Classification Results with C4.5– U series data set

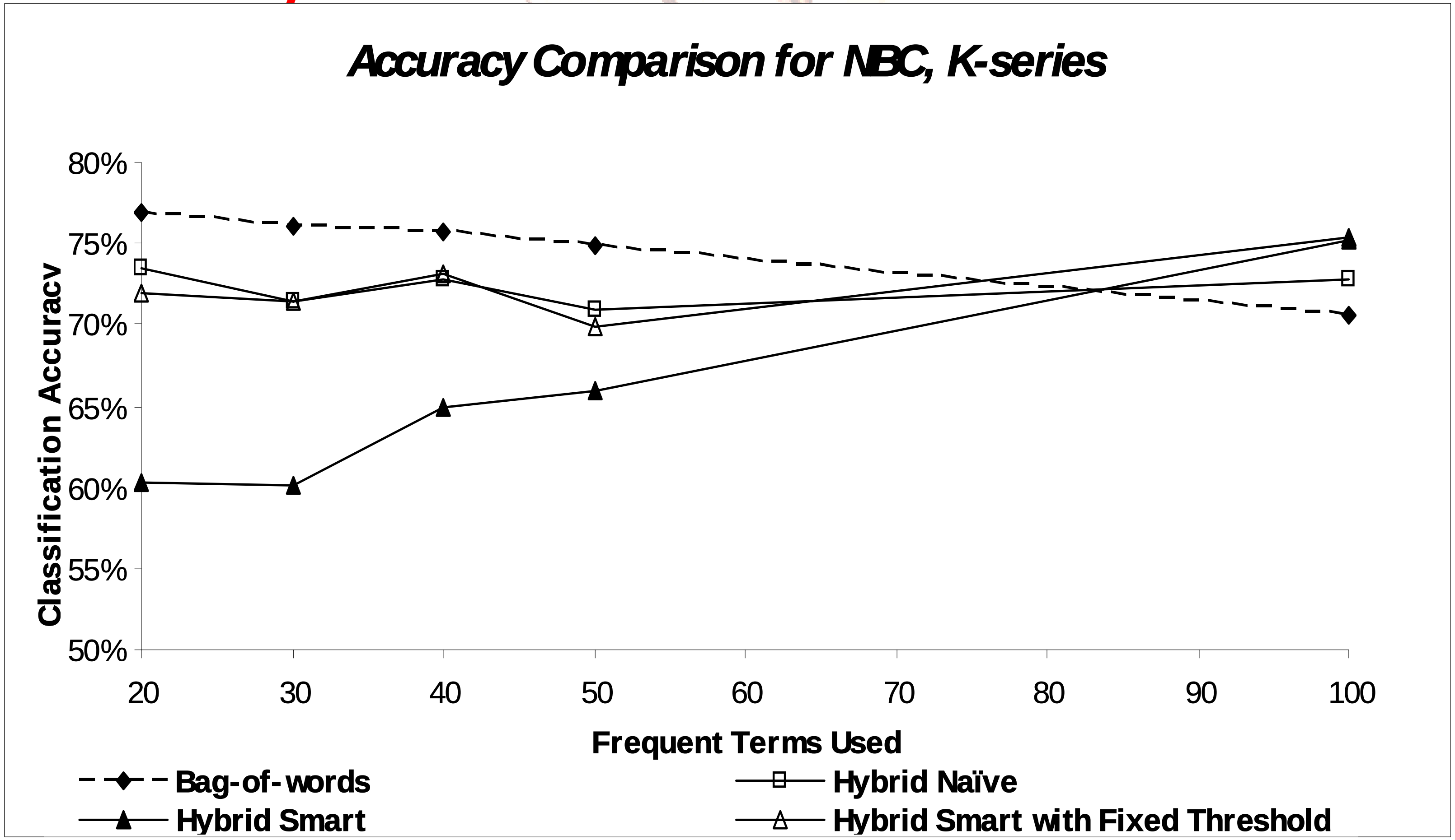


Offline and Online Execution Times for C4.5

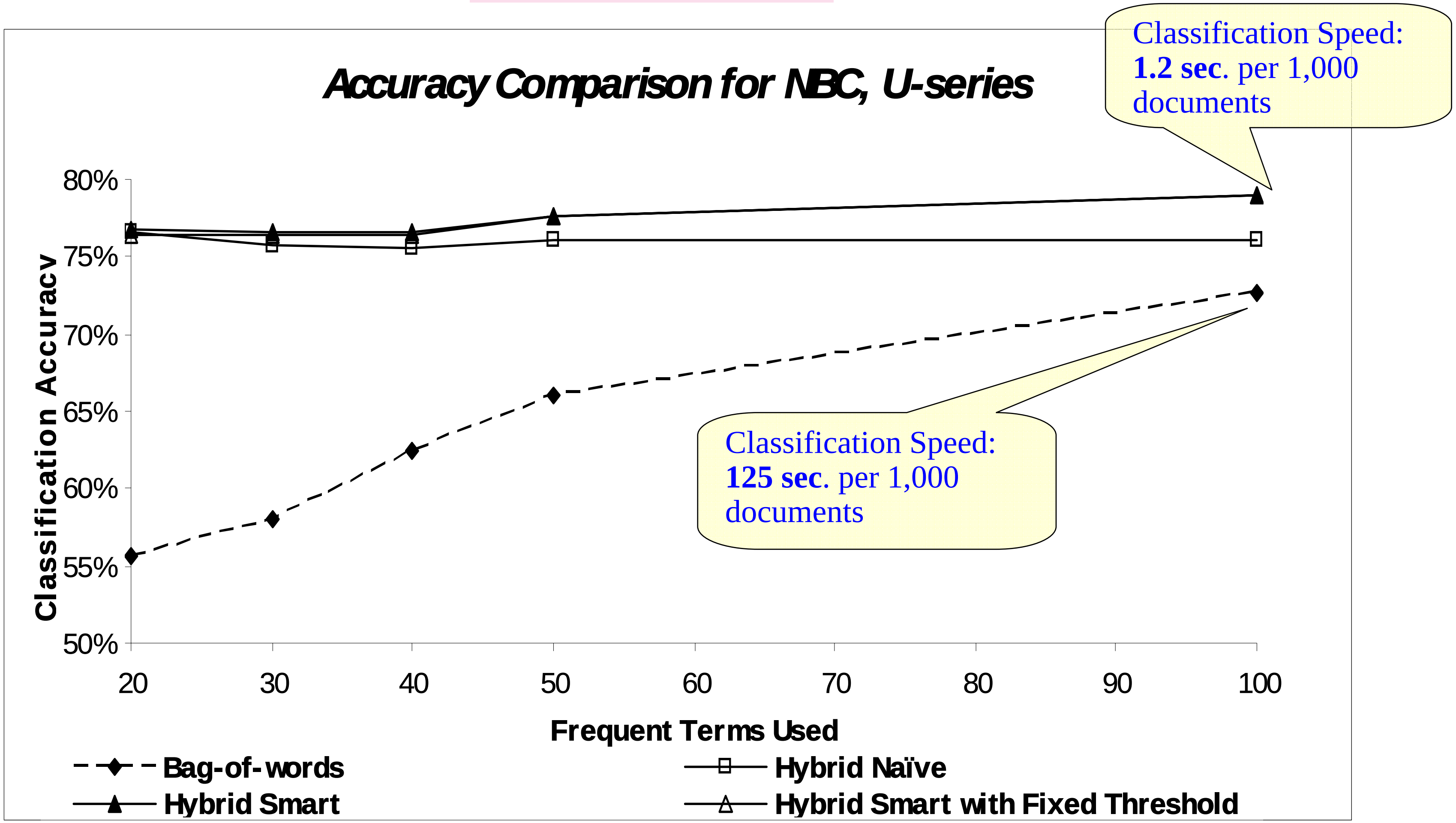
Data Set	Method	Time to Build Graphs (sec)	Time to Build Dictionary (sec)	Time to Construct Vectors (sec)	Time to Build Classification Model (sec)	Total Time Offline (sec)
U-series	Hybrid Smart (N = 100, CR_{min} = 1.1)	223.2	2628.56	5.59	4.36	**2861.71**
	Hybrid Naïve (N = 100, t_{min} = 0.1)	223.2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

Data Set	Method	Average Time to Classify One Document (sec)
U-series	Hybrid Smart	**2.88×10^{-4}**
	Hybrid Naïve	**4.56×10^{-4}**
	Hybrid with Fixed Threshold	**3.12×10^{-4}**
	Bag-of-words	**1.68×10^{-3}**

Classification Results with Naïve Bayes – K series data set



Classification Results with Naïve Bayes – U series data set

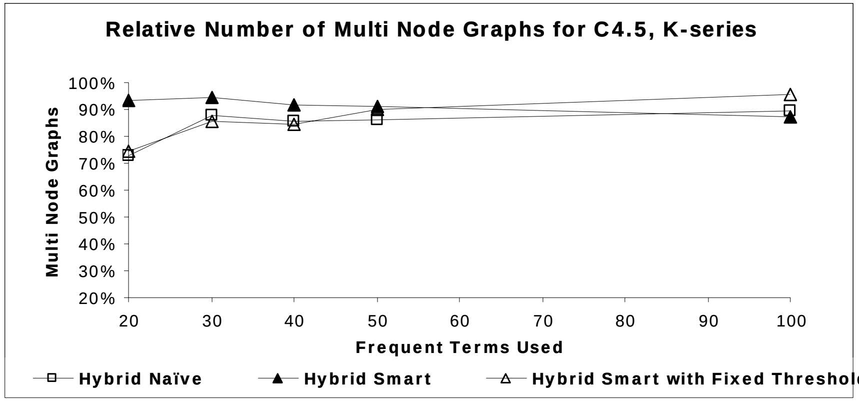


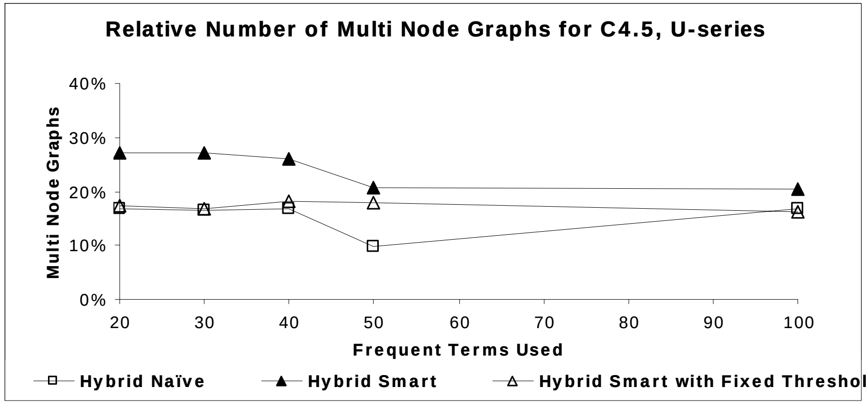
Offline and Online Execution Times for NBC

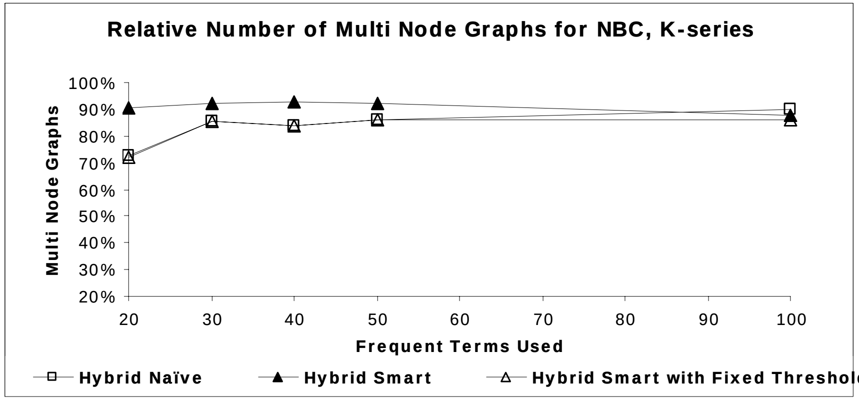
Data Set	Method	Time to Build Graphs (sec)	Time to Build Dictionary (sec)	Time to Construct Vectors (sec)	Time to Build Classification Model (sec)	Total Time Offline (sec)
U-series	Hybrid Smart (N = 100, CR_{min} = 1.2)	223.2	2460.86	4.21	0.12	**2688.4**
	Hybrid Naïve (N = 20, t_{min} = 0.2)	283.64	1.46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

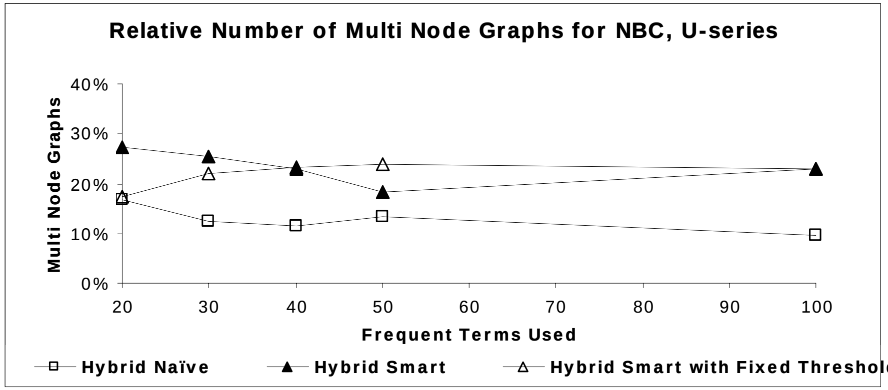
Data Set	Method	Average Time to Classify One Document (sec)
U-series	Hybrid Smart	**1.2×10^{-3}**
	Hybrid Naïve	**6.49×10^{-4}**
	Hybrid with Fixed Threshold	**5.7×10^{-4}**
	Bag-of-words	**0.125**

How many subgraphs have more than one node?









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)

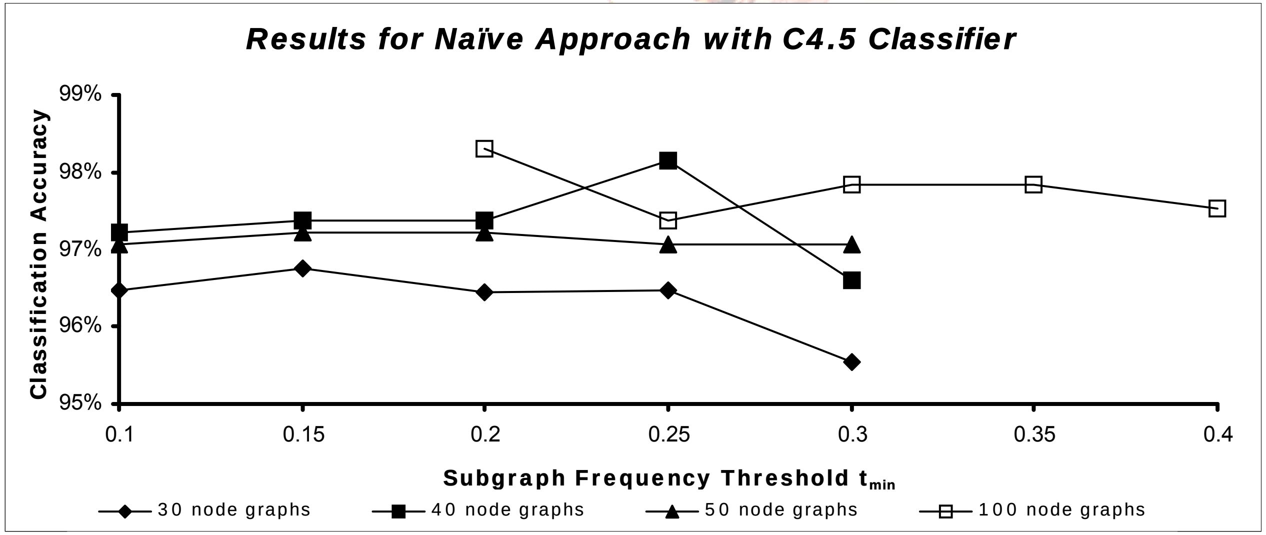
Document Collection

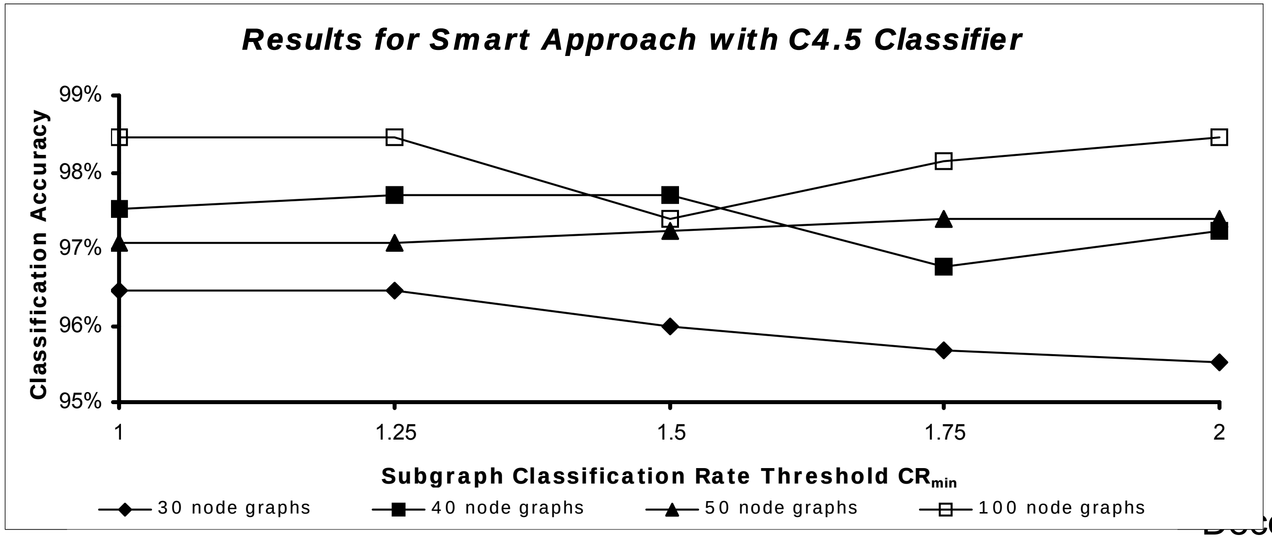
- 648 Arabic documents
  - 200 documents downloaded from terrorist web sites
  - 448 belong to non-terrorist categories
- Terrorist web sites
  - http://www.qudsway.com (Palestinian Islamic Jihad)
  - http://www.palestine-info.com/ (Hamas)
- Normal (non-terrorist) web sites
  - www.aljazeera.net/News
  - http://arabic.cnn.com
  - http://news.bbc.co.uk/hi/arabic/news
  - http://www.un.org/arabic/news

Preprocessing of Documents in Arabic

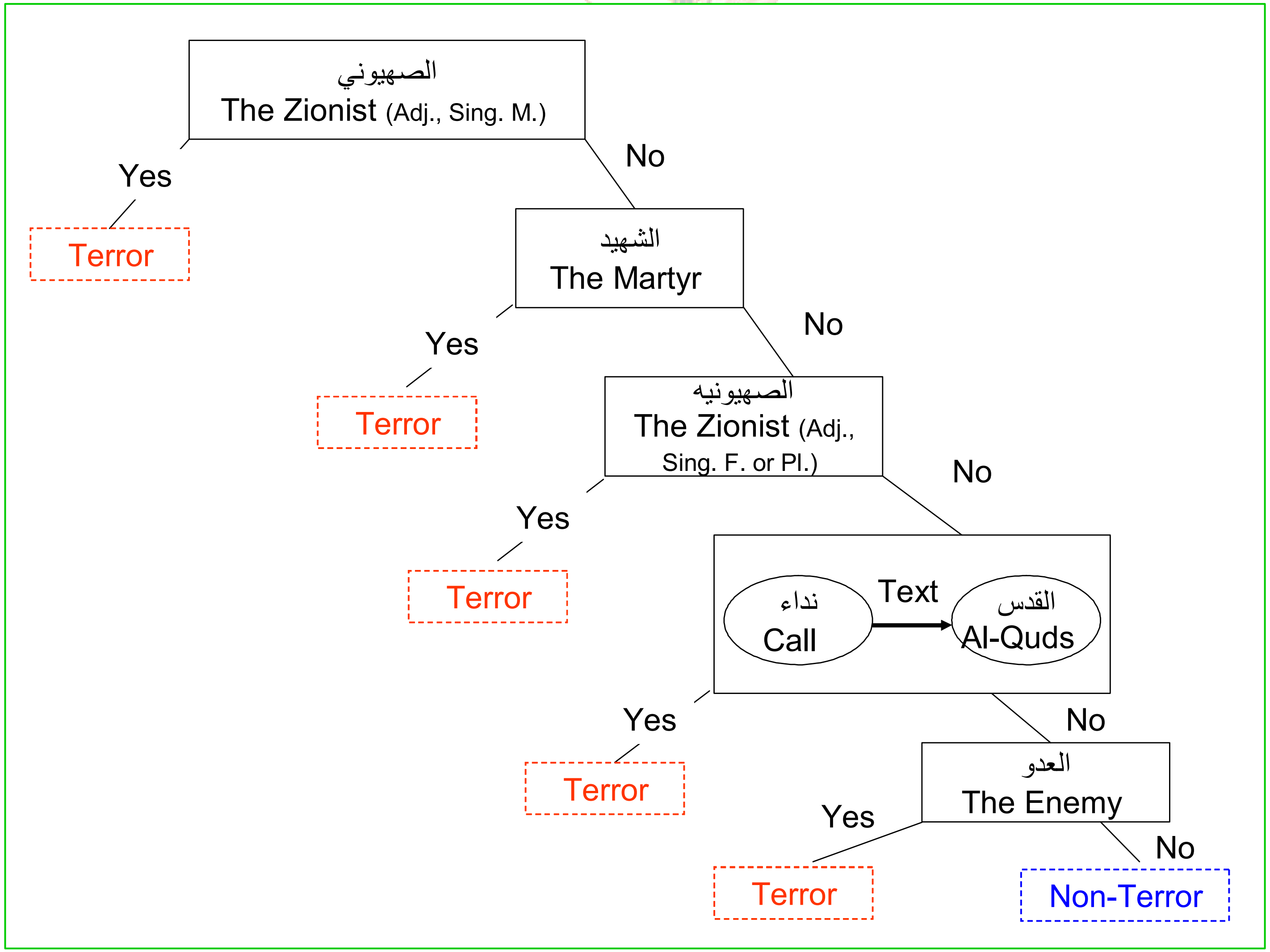
- Normalizing orthographic variations
  - E.g., convert the initial Alif Hamza أ to plain Alif ا
- Normalize the feminine ending, the Ta-Marbuta ة, to Ha ه
- Removal of vowel marks
- Removal of certain letters (such as: Waw و, Kaf ك, Ba ب, and Fa ف) appearing before the Arabic article THE (Alif + Lam ال)
- Removal of pre-defined stop words in Arabic
- Final vocabulary size: 47,836 words

Accuracy Results





Resulting Decision Tree



Does the word الصهيوني ("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:
  - الحركة الصهيونية - The Zionist Movement
  - العدوان الصهيوني – The Zionist enemies
  - المؤامرة الصهيونية – The Zionist plot
  - غلاة الصهيونية - The Zionist extremists
  - المؤتمر الصهيوني الأول – The First Zionist Congress
  - الجماعات الصهيونية المتطرفة – The extremist Zionist groups

Case Study 2

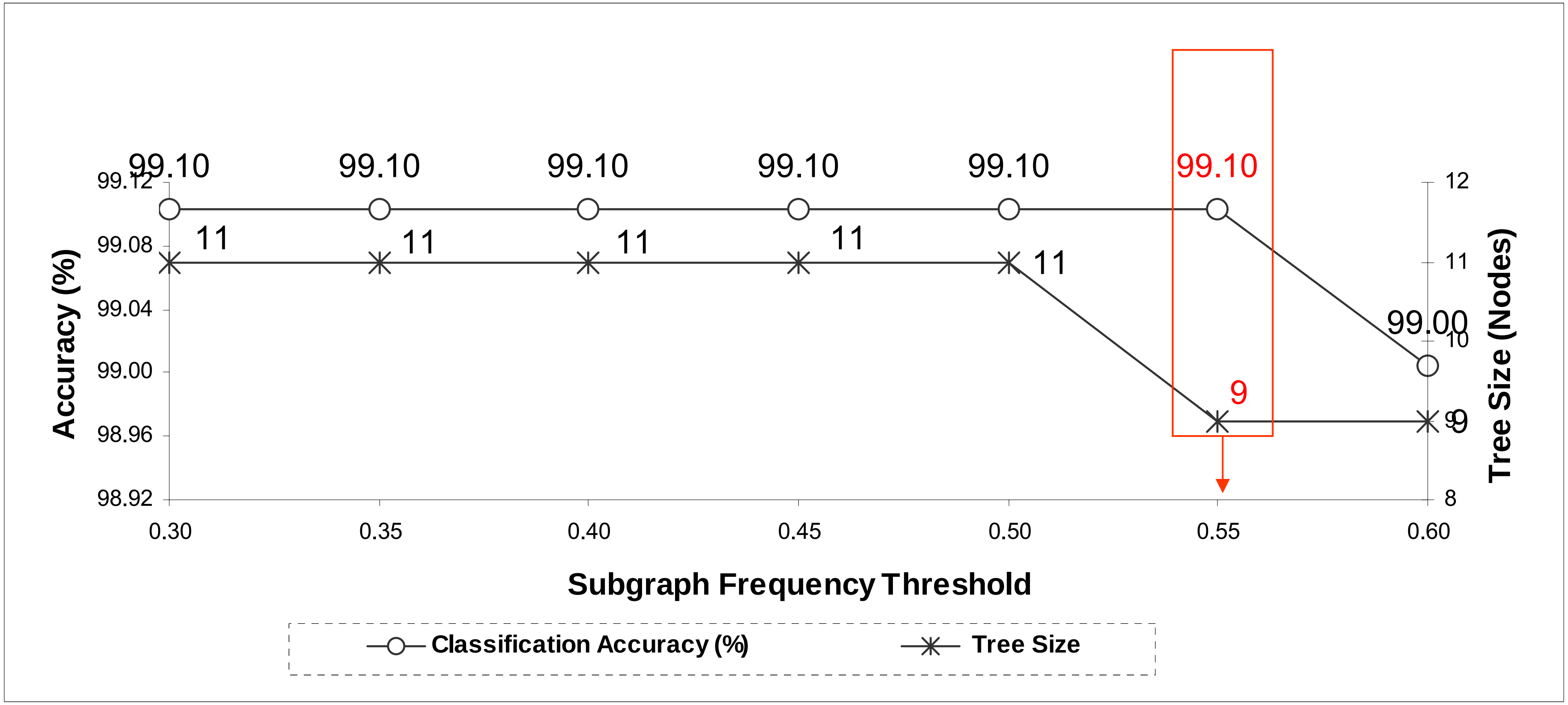
Categorization of Terrorist Web Documents in English

Document Collection

- 1,004 English documents
  - 913 documents downloaded from a Hezbollah web site (http://www.moqawama.org/english/)
  - 91 documents downloaded from a Hamas web site (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

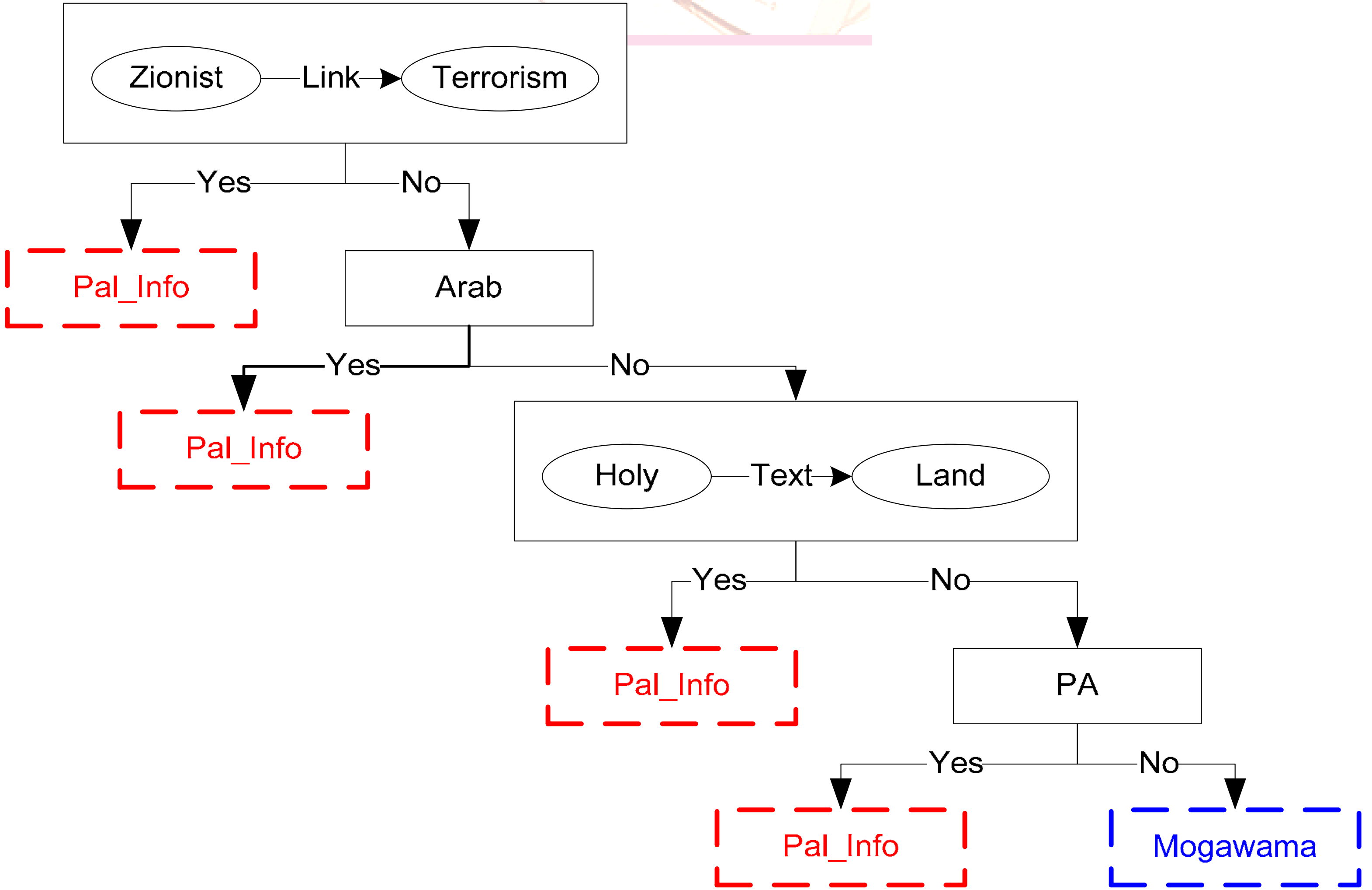
Accuracy Results

Maximum Graph Size: 100 Nodes



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)

- D. Boley, M. Gini, R. Gross, E. H. Han, K. Hastings, G. Karypis, B. Mobasher, J. Moore, “Partitioning-based Clustering for Web Document Categorization”, *Decision Support Systems*, Vol. 27, 1999, pp. 329–341.
- H. Bunke, “On a relation between graph edit distance and maximum common subgraph”, *Pattern Recognition Letters*, Vol. 18, 1997, pp. 689–694.
- H. Bunke and A. Kandel, “Mean and maximum common subgraph of two graphs”, *Pattern Recognition Letters*, Vol. 21, 2000, pp. 163–168.
- H. Bunke and K. Shearer, “A graph distance metric based on the maximal common subgraph”, *Pattern Recognition Letters*, Vol. 19, 1998, pp. 255–259.
- M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam and S. Slattery, "Learning to extract symbolic knowledge from the World Wide Web", In Proceedings of the Fifteenth National Conference on Artificial Intellligence (AAAI98), pages 509-516, 1998.
- M.-L. Fernández and G. Valiente, “A graph distance metric combining maximum common subgraph and minimum common supergraph”, *Pattern Recognition Letters*, Vol. 22, 2001, pp. 753–758.
- T. Joachims, “Learning to Classify Text Using Support Vector Machines, Methods, Theory and Algorithms”, Kluwer, 2002.

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).
- M. Last, "Using Data Mining Technology for Terrorist Detection on the Web", in M. Last and A. Kandel (Editors), Fighting Terror in Cyberspace, World Scientific, Series in Machine Perception and Artificial Intelligence, Vol. 65, pp. 41-62, 2005.
- M. Last, A. Markov, and A. Kandel, "Multi-Lingual Detection of Terrorist Content on the Web", Proceedings of the PAKDD'06 International Workshop on Intelligence and Security Informatics (WISI'06), Lecture Notes in Computer Science, Vol. 3917, pp. 16-30, Springer, 2006.
- A. Markov and M. Last, "Identification of Terrorist Web Sites with Cross-Lingual Classification Tools", in M. Last and A. Kandel (Editors), Fighting Terror in Cyberspace, World Scientific, Series in Machine Perception and Artificial Intelligence, Vol. 65, pp. 117-141, 2005.
- A. Markov and M. Last, "Efficient Graph-Based Representation of Web Documents", Proceedings of the Third International Workshop on Mining Graphs, Trees and Sequences (MGTS2005), pp. 52-62, October 7, 2005, Porto, Portugal.

References (3)

- A. Markov, M. Last, and A. Kandel, “Model-Based Classification of Web Documents Represented by Graphs”, Proceedings of WebKDD 2006 Workshop on Knowledge Discovery on the Web at KDD 2006, pp. 31-38, Philadelphia, PA, USA, Aug. 20, 2006.
- G. Salton, A. Wong, and C. Yang, C. (1975). A Vector Space Model for Automatic Indexing, Comm. of the ACM, 18(11), pp. 613--620.
- G. Salton, and M. McGill, "Introduction to Modern Information Retrieval", McGraw Hill, 1983.
- A. Schenker, H. Bunke, M. Last, A. Kandel, "Graph-Theoretic Techniques for Web Content Mining", World Scientific, 2005.
- A. Schenker, M. Last, H. Bunke, A. Kandel, "Classification of Web Documents Using Graph Matching", International Journal of Pattern Recognition and Artificial Intelligence, Vol. 18, No. 3, pp. 475-496, 2004.
- P.D. Turney, "Learning Algorithms for Keyphrase Extraction," *Information Retrieval*, 2 (4), pp. 303-336, 2000.
- W. D. Wallis, P. Shoubridge, M. Kraetz, and D. Ray, “Graph distances using graph union”, *Pattern Recognition Letters*, Vol. 22, 2001, pp. 701–704.