Graph-Based Methods for Multilingual Text and Web Mining

Mark Last

Department of Information Systems Engineering

Ben-Gurion University of the Negev

In cooperation with

Horst Bunke (University of Bern)

Abraham Kandel, Adam Schenker (University of South Florida)

Alex Markov, Marina Litvak, Guy Danon (Ben-Gurion University)

E-mail: mlast@bgu.ac.il

Home Page: http://www.bgu.ac.il/~mlast/

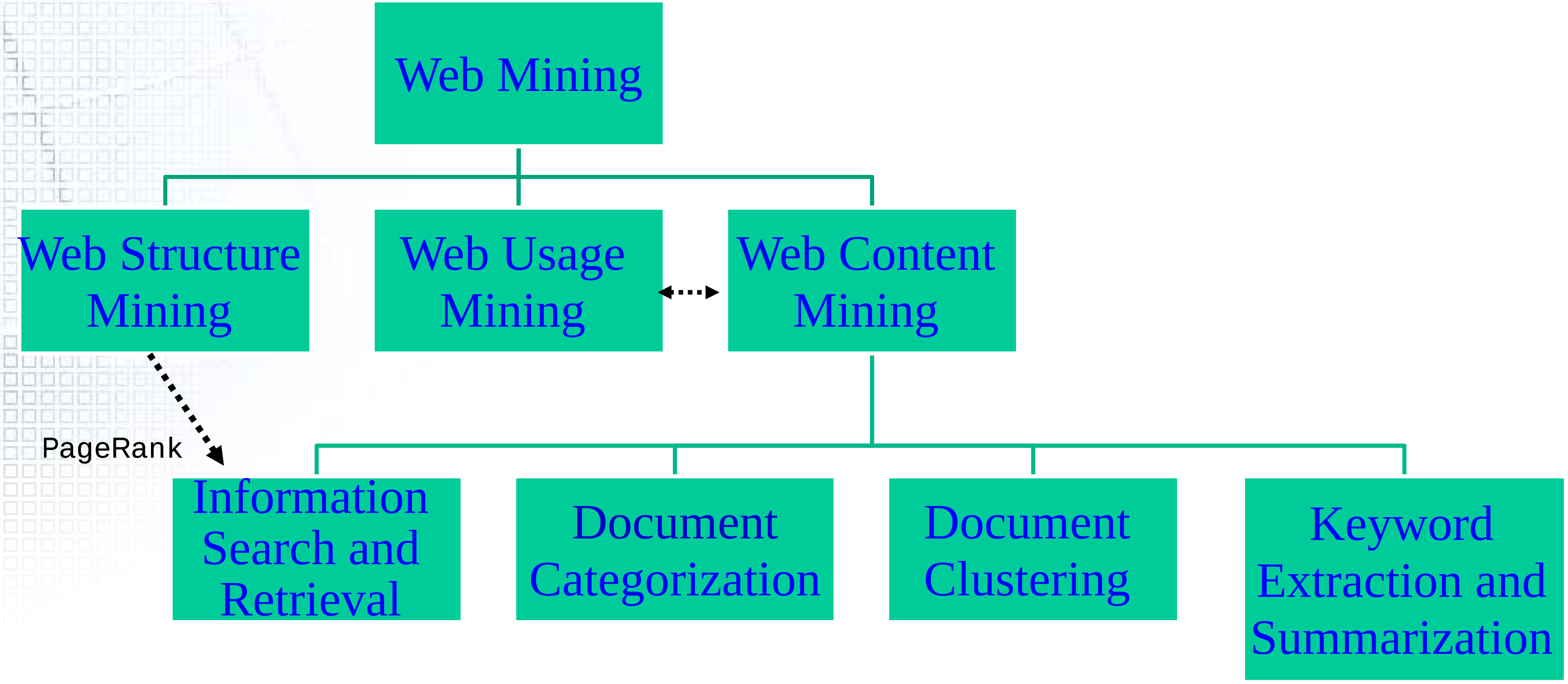
Text Mining Day 2009 at BGU, May 25, 2009

Agenda

- Introduction and Motivation
- Graph-Based Representations of Text and Web Documents
- Graph-Based Categorization and Clustering Algorithms
- The Hybrid Approach to Web Document Categorization
- Graph-Based Keyword Extraction
- Summary

INTRODUCTION AND MOTIVATION

Web Mining Tasks



Web Mining

Web Structure Mining

Web Usage Mining

Web Content Mining

PageRank

Information Search and Retrieval

Document Categorization

Document Clustering

Keyword Extraction and Summarization

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

Advantages of the Vector-Space Model

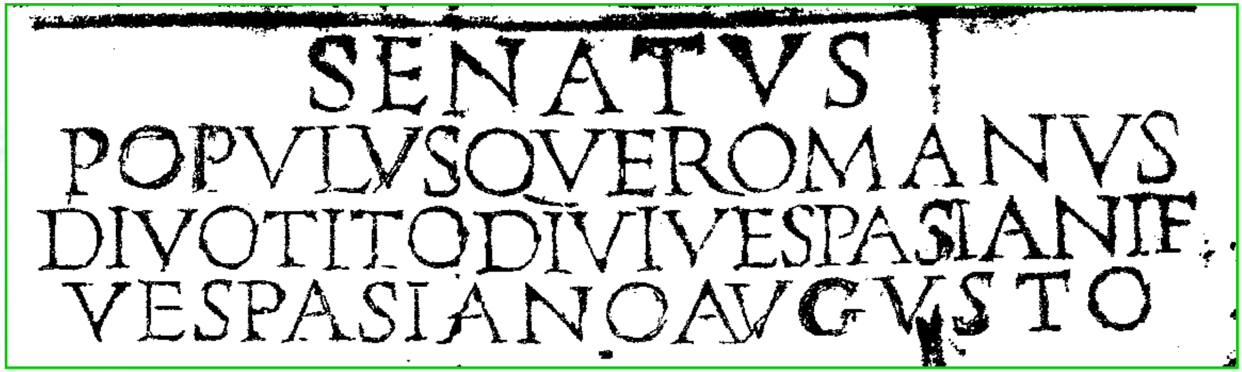
(based on Joachims, 2002)

- 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)

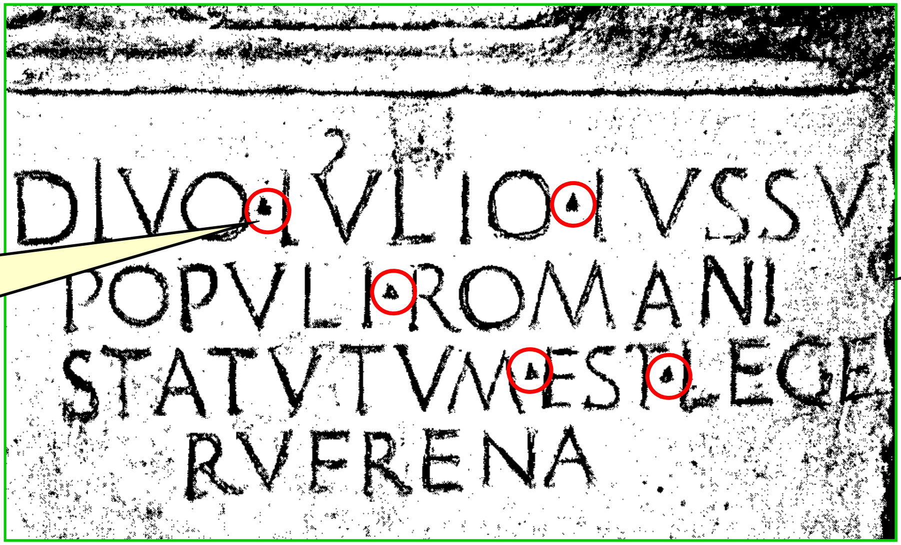
Limitations of the Vector-Space Model

- 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

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

Introduced in Schenker *et al*., 2005

GRAPH-BASED REPRESENTATIONS OF TEXT AND WEB DOCUMENTS

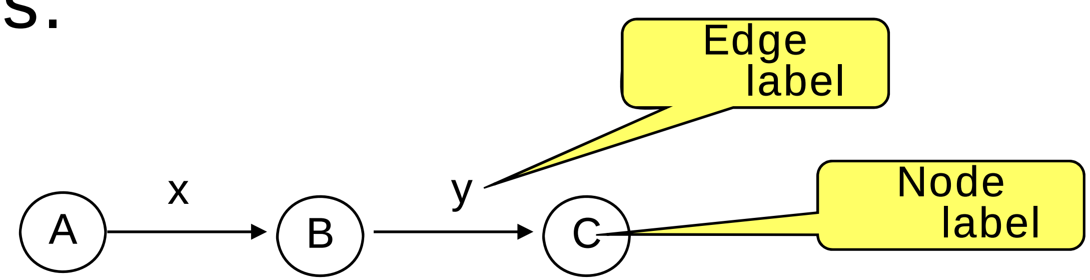
Relevant Definitions

(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.



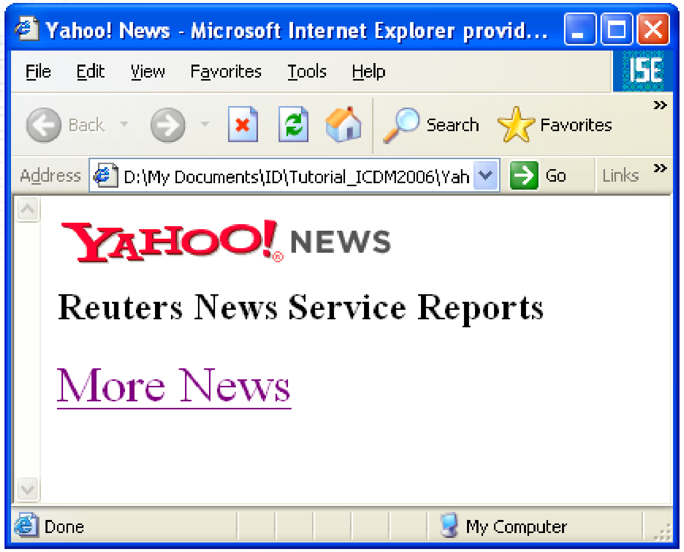
- Node and edge IDs are omitted for brevity
- **Graph size**: $|G|=|V|+|E|$

The Graph-Based Model of Web Documents – Basic Ideas

- At most one node for each unique term in a document
- If a word *B* follows a word *A*, there is a directed edge from *A* to *B*
  - Unless the words are separated by certain punctuation marks (periods, question marks, and exclamation points)
- Stop words are removed
- Graph size may be 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

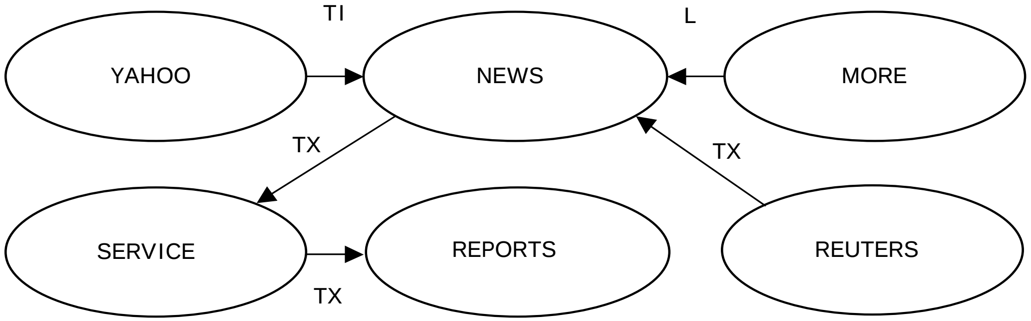
- Edges are labeled according to the document section where the words are followed by each other
  - *Title (TI)* contains the text related to the document's title and any provided keywords (meta-data);
  - *Link (L)* is the "anchor text" that appears in clickable hyper-links on the document;
  - *Text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


Yahoo! News - Microsoft Internet Explorer provid...
File Edit View Favorites Tools Help
Back Search Favorites
Address D:\My Documents\ID\Tutorial_ICDM2006\Yah Go Links
Done
My Computer

YAHOO! NEWS

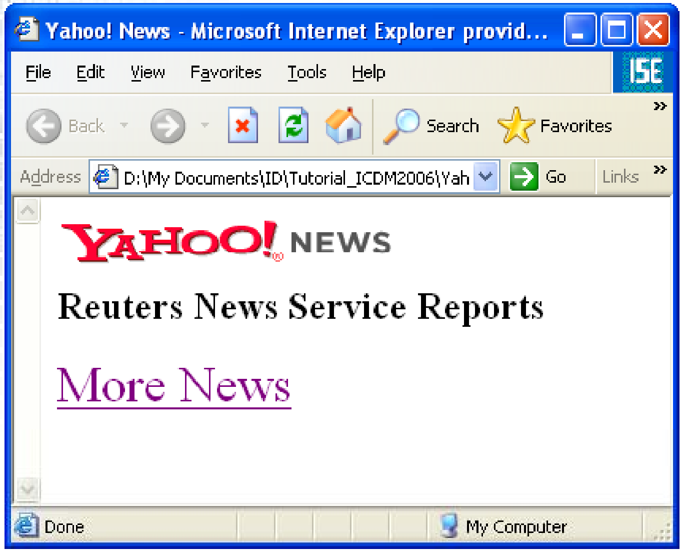
Reuters News Service Reports

More News



The Simple Representation

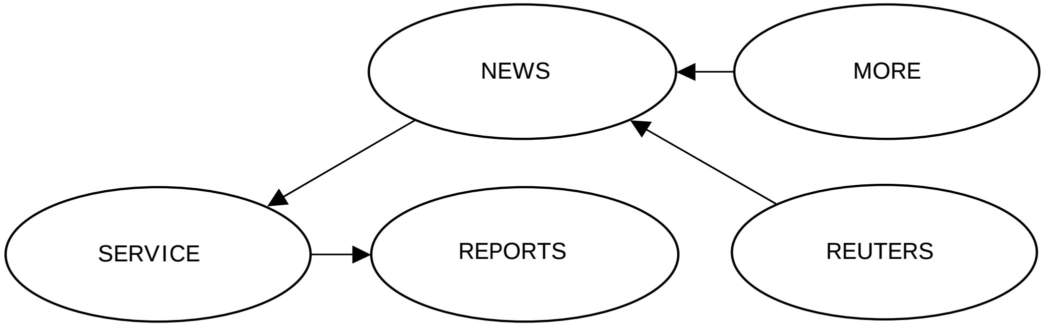
- The graph is based only the visible text on the page (title and meta-data are ignored)
- Edges are not labeled


Yahoo! News - Microsoft Internet Explorer provid...
File Edit View Favorites Tools Help
Back
Search
Favorites
Address
D:\My Documents\ID\Tutorial_ICDM2006\Yah
Go
Links
Done
My Computer

YAHOO! NEWS

Reuters News Service Reports

More News

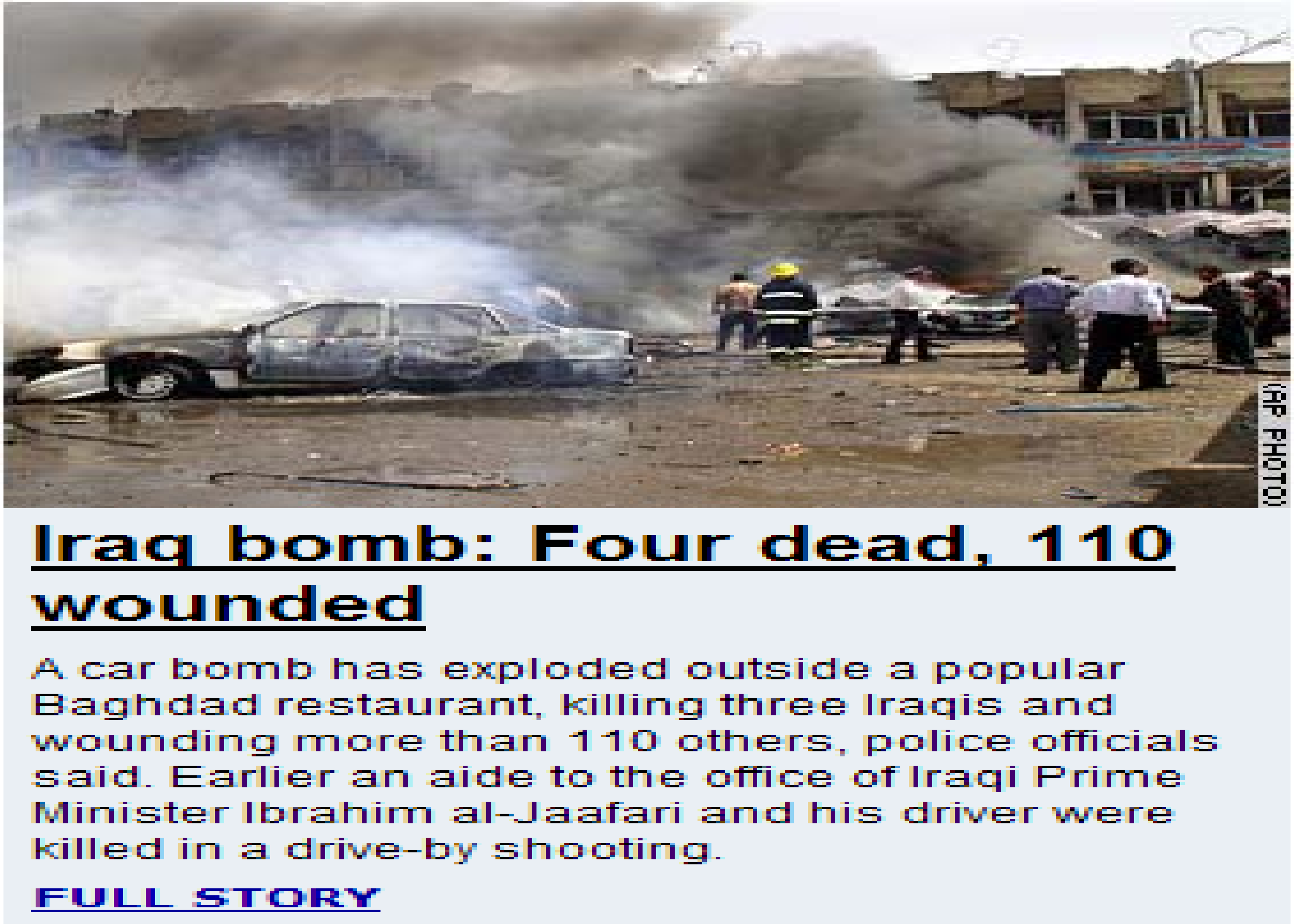


Other Representations

- The *n*-distance Representation
  - Look up to *n* terms ahead and connect the succeeding terms with an edge that is labeled with the distance between them (*n*)
- The *n*-simple Representation
  - Look up to *n* terms ahead and connect the succeeding terms with an unlabeled edge
- The Absolute Frequency Representation
  - Each node and edge is labeled with an absolute frequency measure
- The Relative Frequency Representation
  - Each node and edge is labeled with a relative frequency measure

Graph Based Document Representation

Example –Source: www.cnn.com, 24/05/2005

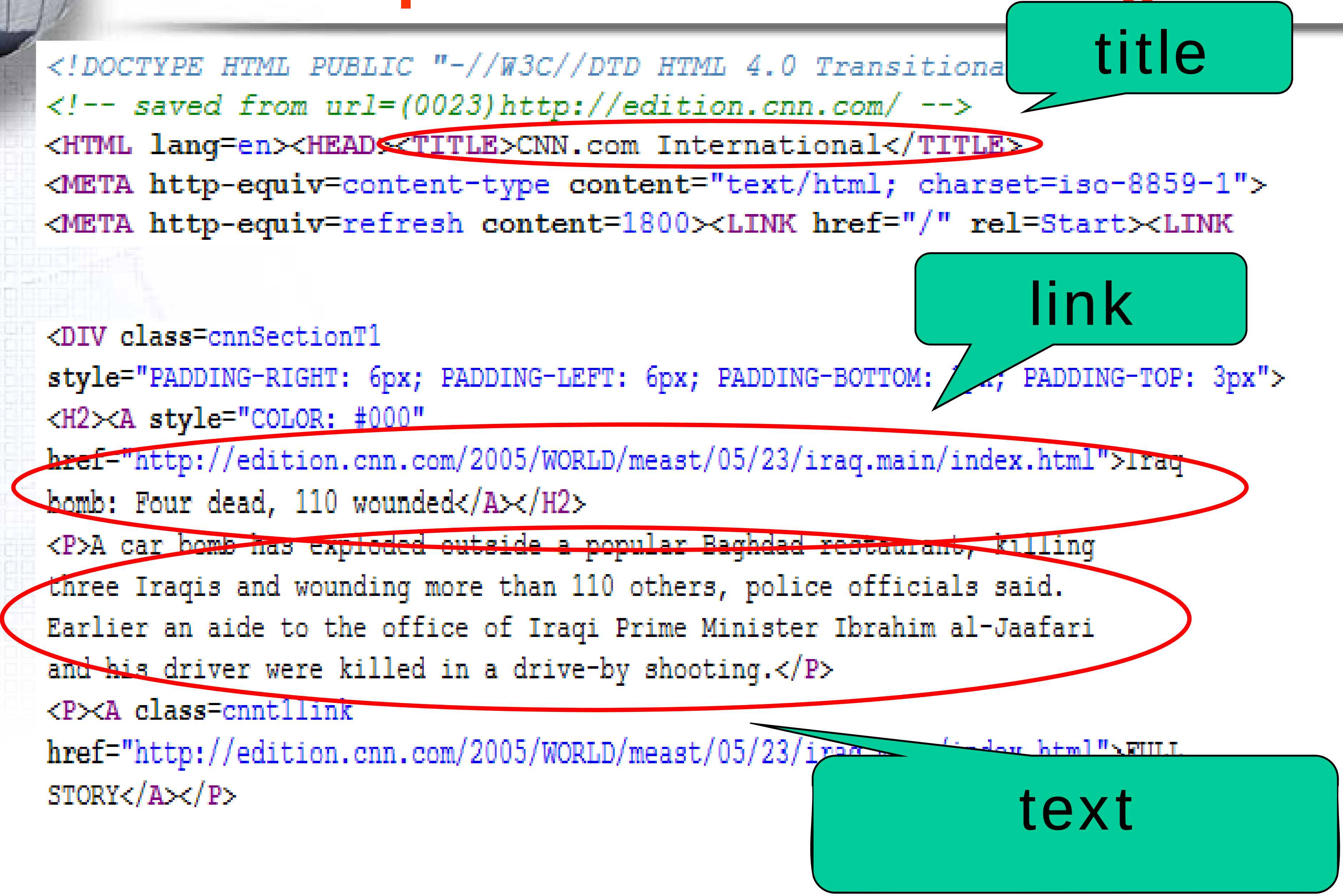


Iraq bomb: Four dead, 110 wounded

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 Transitiona
<!-- 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: 1px; 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>
```

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 wound .

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

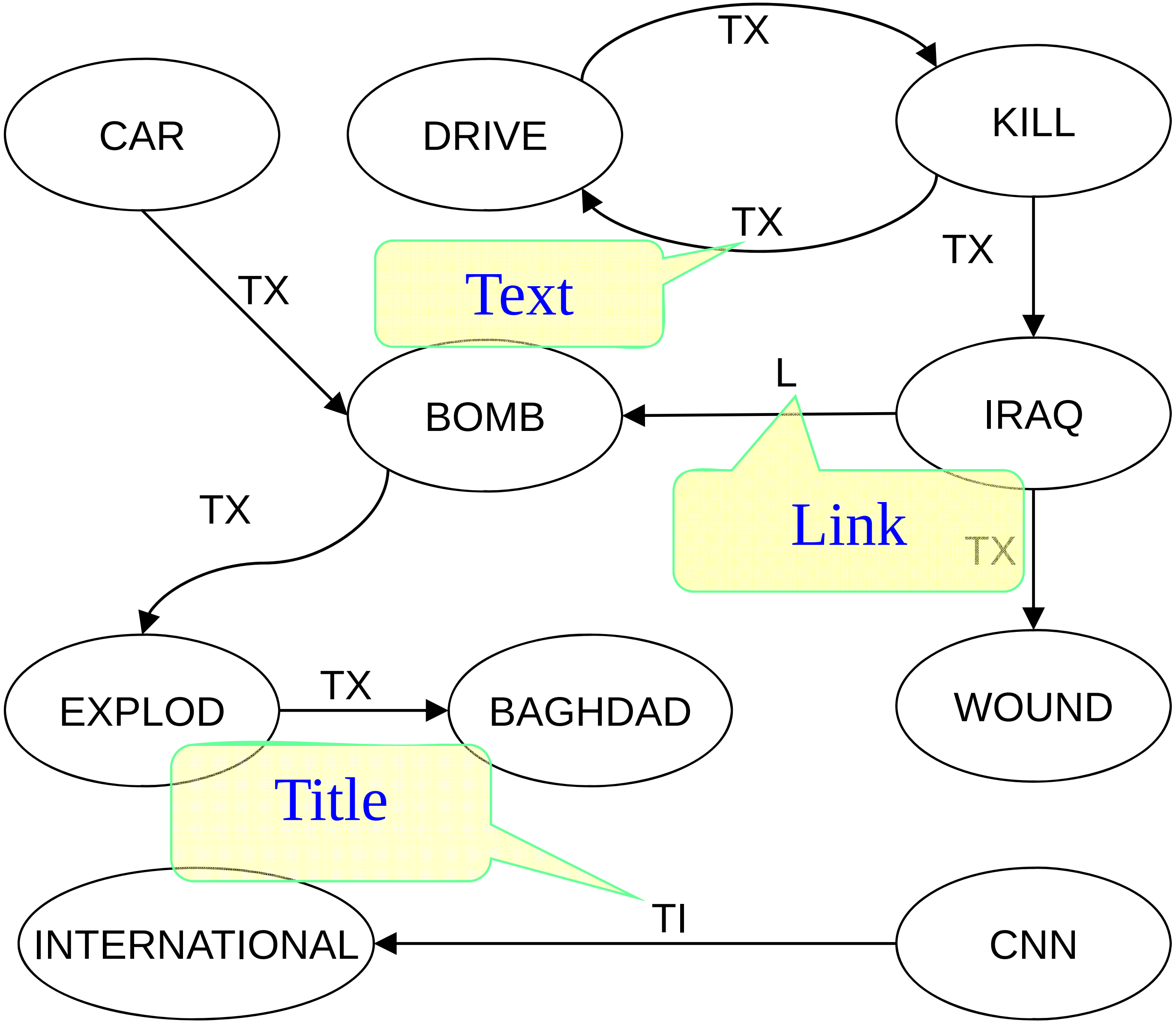
Iraqis bomb: Four dead, 110 wounding.

FULL STORY.

Standard Graph Based Document Representation

Ten most frequent terms are used

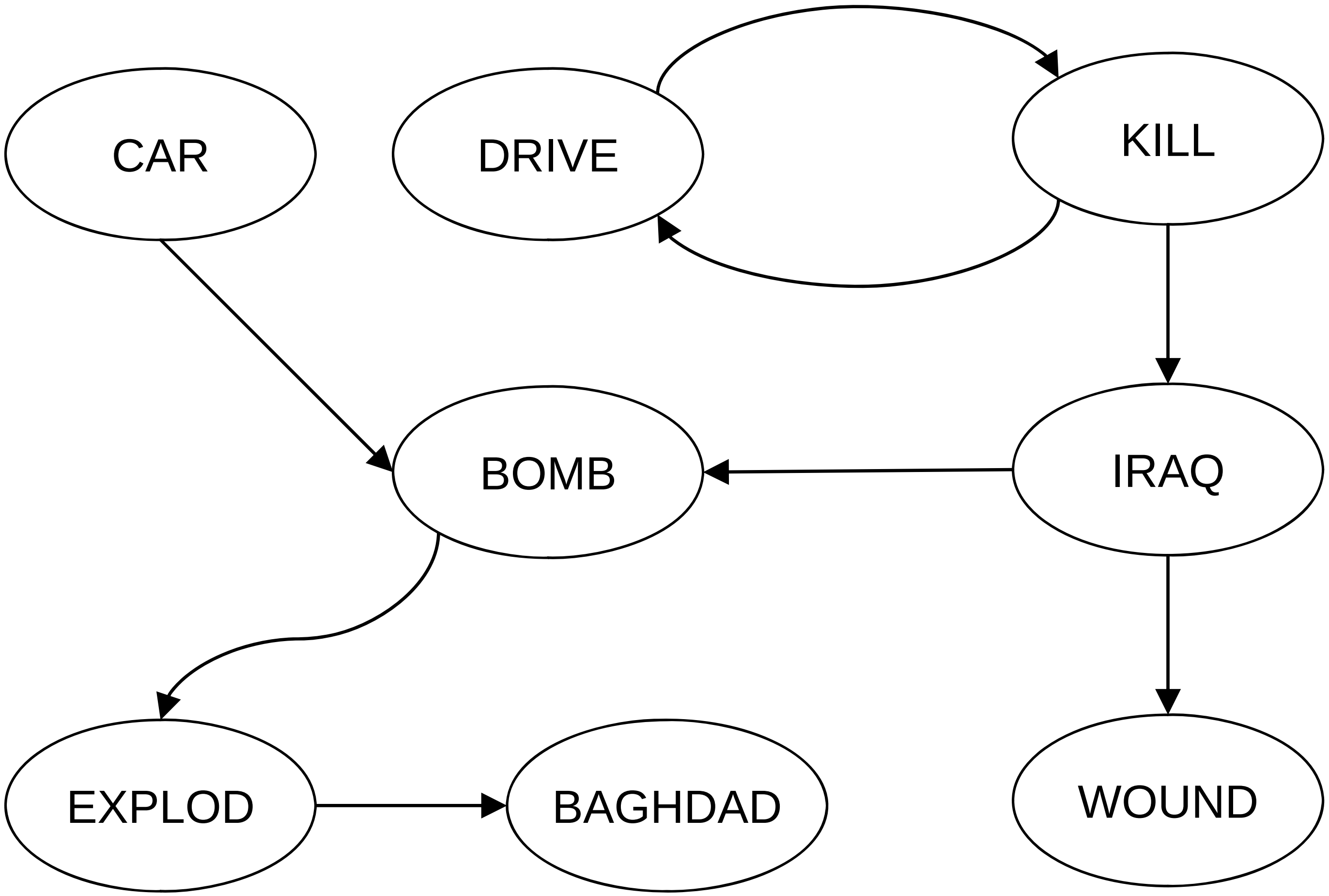
Word	Frequency
Iraq	3
Kill	2
Bomb	2
Wound	2
Drive	2
Explod	1
Baghdad	1
International	1
CNN	1
Car	1



Simple Graph Based Document Representation

Ten most frequent terms are used

Word	Frequency
Iraq	3
Kill	2
Bomb	2
Wound	2
Drive	2
Explod	1
Baghdad	1
International	1
CNN	1
Car	1



Based on Schenker *et al*., 2005

GRAPH-BASED CATEGORIZATION AND CLUSTERING ALGORITHMS

"Lazy" Document Categorization with Graph-Based Models

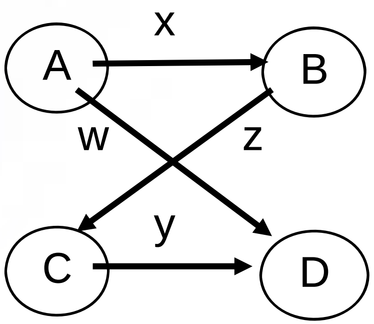
- The Basic *k*-Nearest Neighbors (*k*-NN) Algorithm
  - *Input*: a set of labeled training documents, a query document *d*, and a parameter *k* defining the number of nearest neighbors to use
  - *Output*: a label indicating the category of the query document *d*
  - *Step 1*. Find the *k* nearest training 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
  - Use a graph-theoretical **distance measure**

Distance between two Graphs

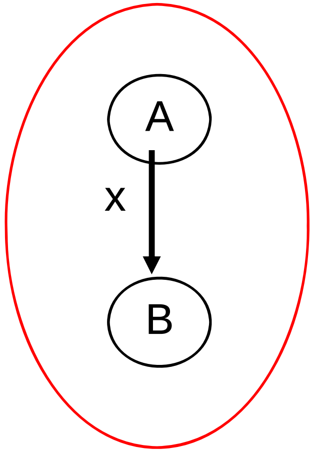
- Required properties
  - (1) *boundary condition*: $d(G_1,G_2)\geq 0$
  - (2) *identical graphs have zero distance*: $d(G_1,G_2)=0 \rightarrow G_1 \cong G_2$
  - (3) *symmetry*: $d(G_1,G_2)=d(G_2,G_1)$
  - (4) *triangle inequality*: $d(G_1,G_3)\leq d(G_1,G_2)+d(G_2,G_3)$

Maximum Common Subgraph (mcs)

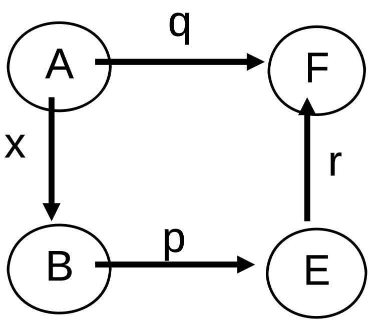
- The graph G is a **maximum common subgraph (mcs)** if G is a common subgraph of G_1 and G_2 and there exist no other common subgraph G' of G_1 and G_2 such that $|G'| > |G|$



$\mathbf{G_1}$



G

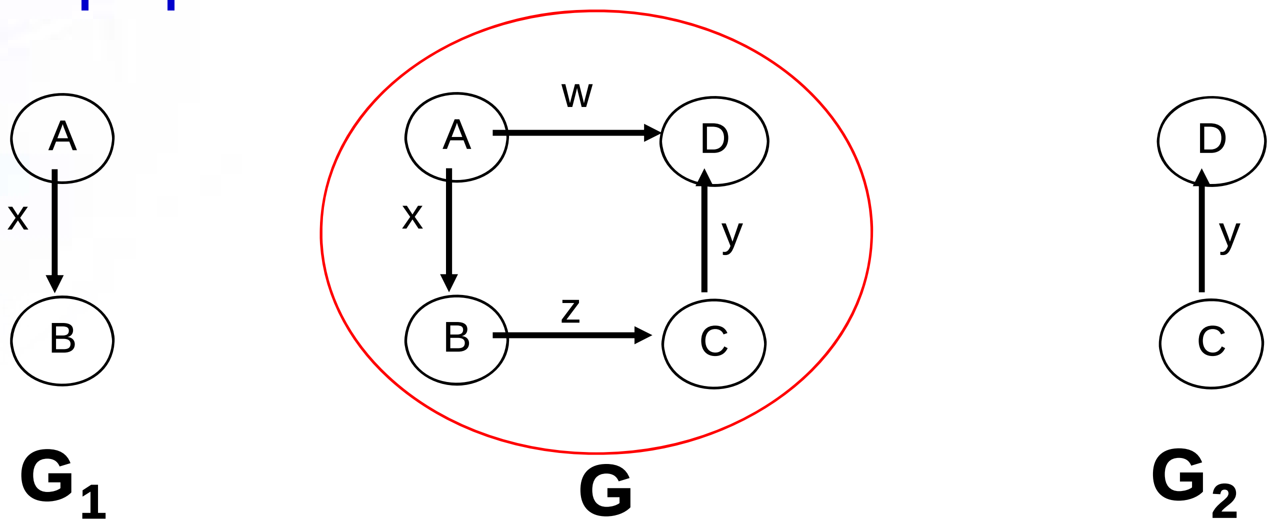


$\mathbf{G_2}$

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

Minimum Common Supergraph (MCS)

- 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 $|G'| < |G|$



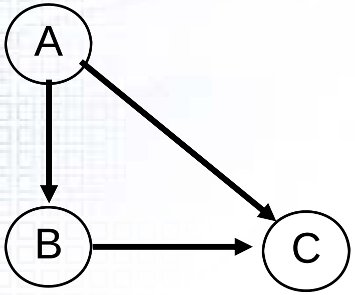
$$|G| = |V| + |E| = 4 + 2 = 6$$

MMCSN Distance between two Graphs

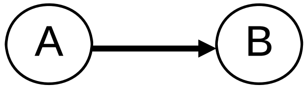
- MMCSN Measure (Schenker et al., 2005):

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

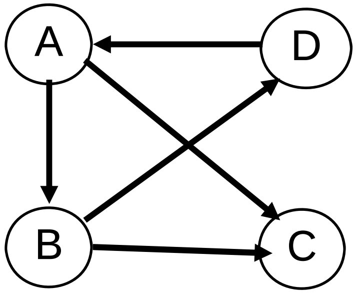
- *mcs(G_1, G_2)* - maximum common subgraph
- *MCS(G_1, G_2)* - minimum common supergraph



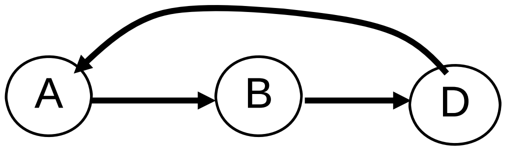
G_1



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):

$$d_{UGU}(G_1,G_2) = |G_1| + |G_2| - 2|mcs(G_1,G_2)|$$

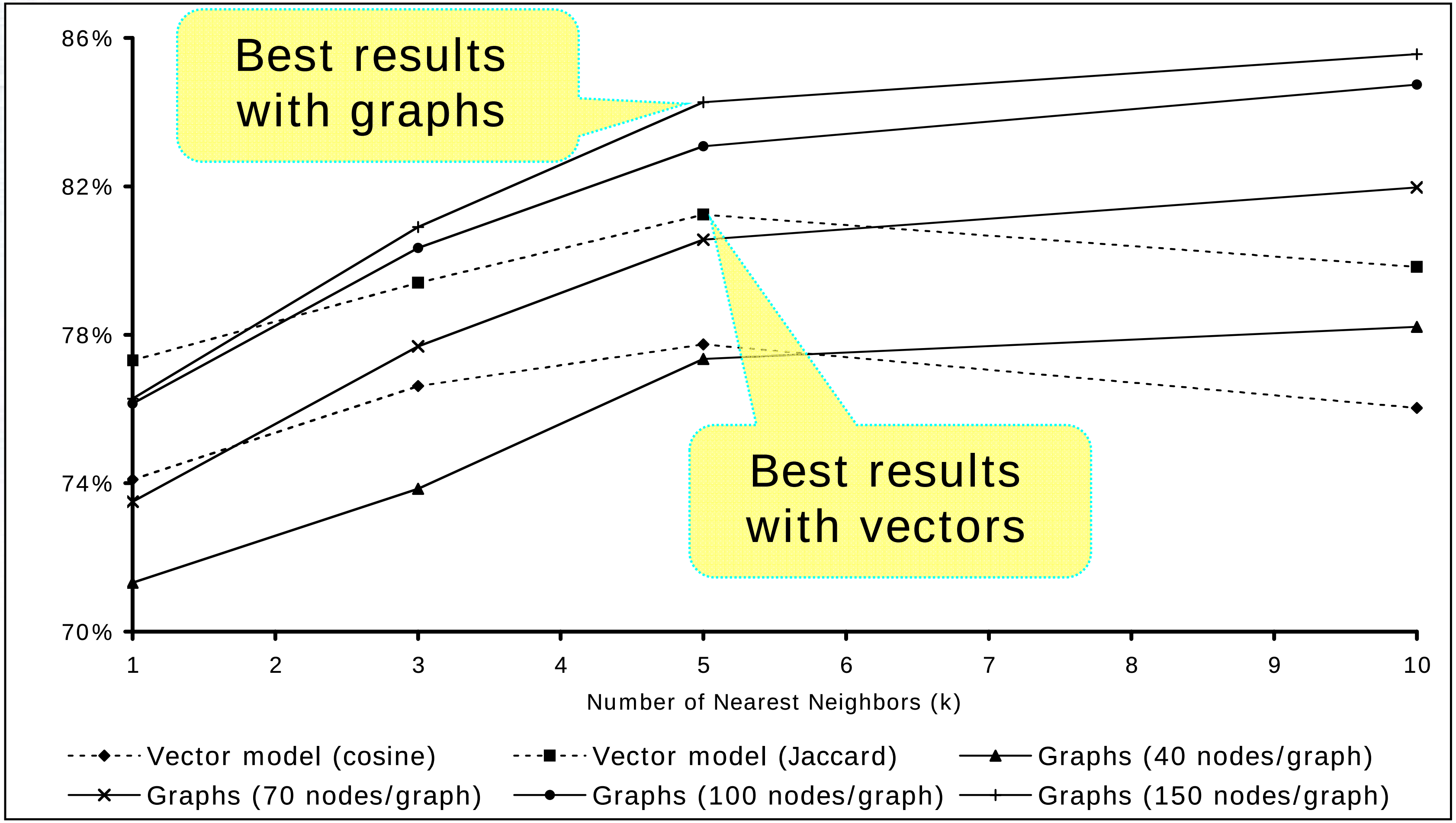
- Fernández and Valiente (2001):

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

k-Nearest Neighbors with Graphs

Sample Accuracy Results (Schenker et al., 2004)

Benchmark Data Set: K-series (Boley *et al*., 1999)
2,340 web documents from 20 categories
Source: English news pages hosted at Yahoo!



Best results with graphs

Best results with vectors

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

“Lazy” Document Categorization with Graph-Based Models

- Advantages
  - Keeps HTML structure information
  - Retains original order of words
  - Outperforms the vector-space model with several distance measures
- Limitation
  - Can work only with “lazy” classifiers (such as *k*-NN), which have a very low classification speed
- Conclusion
  - Graph models cannot be used directly for fast, model-based classification of web documents (e.g., using a decision tree)
- Solution
  - The **hybrid approach**: represent a document as a vector of sub-graphs (in a few minutes…)

The Graph-Based *k*-Means Clustering Algorithm

Inputs: the set of *n* data items (represented by graphs) and a parameter *k*, defining the number of clusters to create

Outputs: the centroids of the clusters (represented by median graphs) and for each data item the cluster (an integer in [1,*k*]) it belongs to

Step 1. Assign each data item randomly to a cluster (from 1 to *k*).

Step 2. Using the initial assignment, determine the median of the set of graphs of each cluster.

Step 3. Given the new medians, assign each data item to be in the cluster of its closest median, using a graph-theoretic distance measure.

Step 4. Re-compute the medians as in Step 2. Repeat Steps 3 and 4 until the medians do not change.

Median of a set of graphs S (Bunke et al., 2001) is a graph *g*∈*S* such that *g* has the lowest average distance to all elements in *S*:

$$g = \arg\min_{\forall s \in S} \left(\frac{1}{|S|} \sum_{i=1}^{|S|} d(s, G_i) \right)$$

Graph-Based Document Clustering

Comparative Evaluation – Dunn Index

$$D_I = \frac{d_{\min}}{d_{\max}}$$

$d_{\min}$ - the minimum distance between any two objects in different clusters

$d_{\max}$ - the maximum distance between any two items in the same cluster

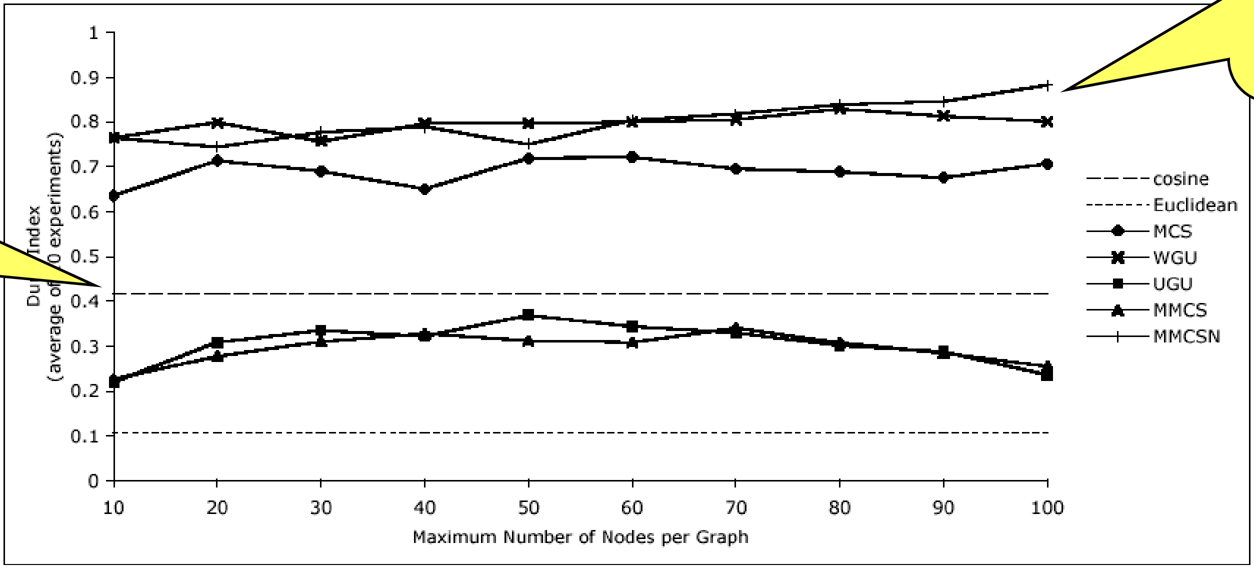


Figure 7.3. Distance Measure Comparison for the F-Series Data Set (Dunn Index)

The best graph-based methods

The best vector-based method

Presented in Markov *et al*., 2008

THE HYBRID APPROACH TO WEB DOCUMENT CATEGORIZATION

The Hybrid Approach to Document Categorization

(Markov *et al*., 2006)

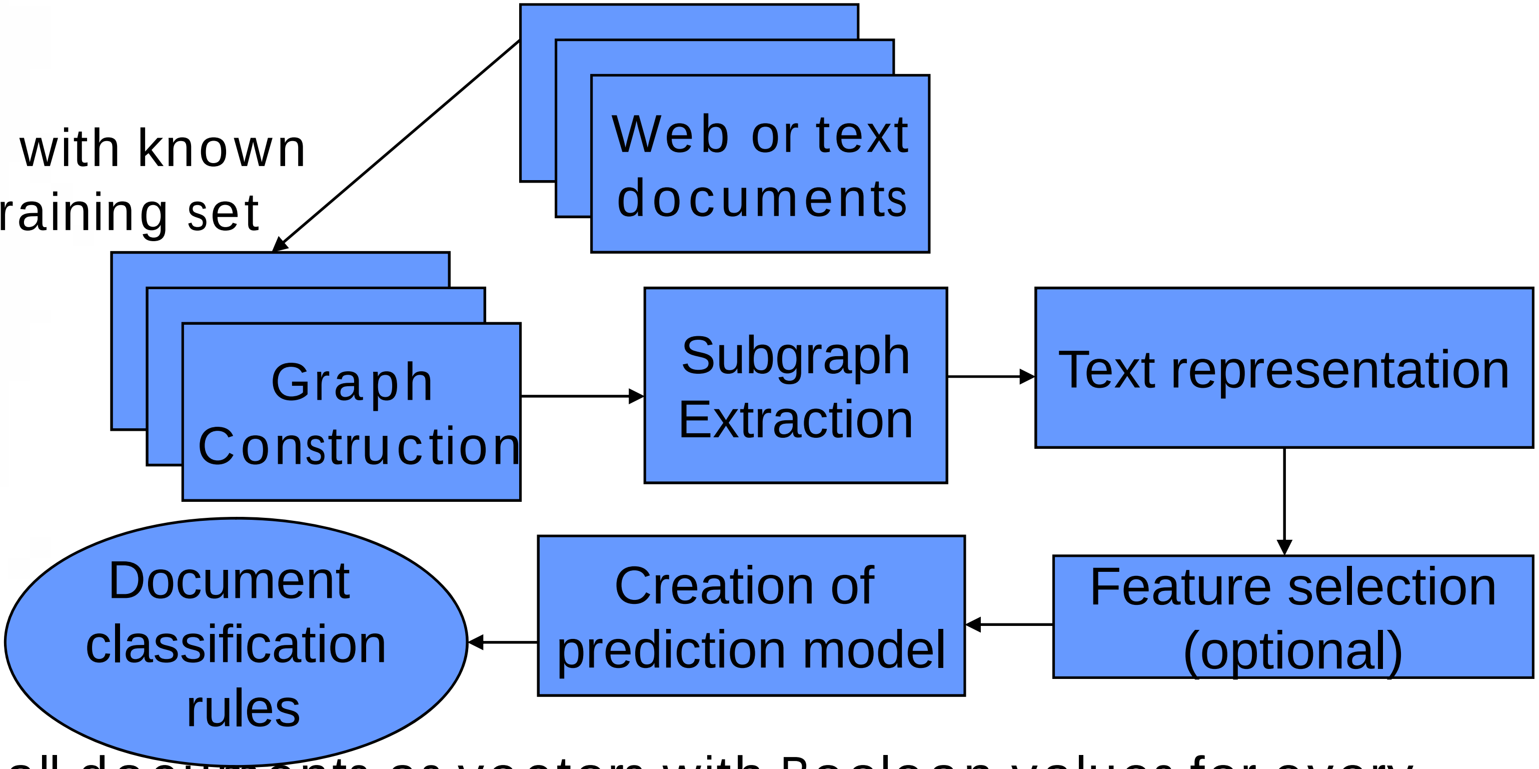
- Basic Idea
  - Represent a document as a vector of sub-graphs
  - Categorize documents with a *model-based classifier* (e.g., a decision tree), which is much faster than a “lazy” method
- Naïve Approach
  - Select sub-graphs that are most frequent in each category
- Smart Approach
  - Select sub-graphs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

Predictive Model Induction with Hybrid Representation (Markov *et al*., 2006)

Set of documents with known categories – the training set

Documents graph representation

Extraction of sub-graphs ***relevant for classification***



Web or text documents
Graph Construction
Subgraph Extraction
Text representation
Feature selection (optional)
Creation of prediction model
Document classification rules

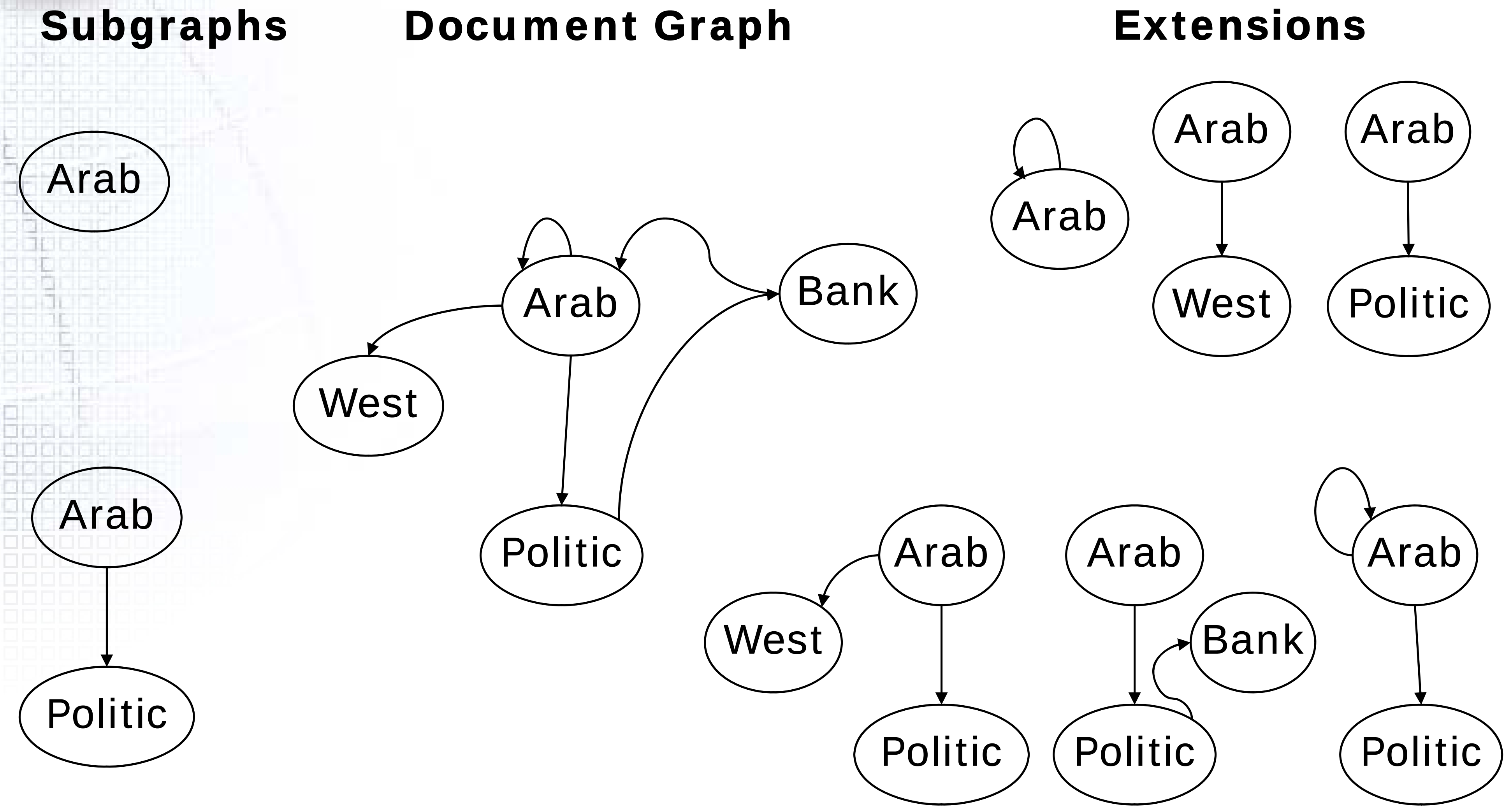
Representation of all documents as vectors with Boolean values for every sub-graph in the set

Identification of best attributes (Boolean features) for classification

Finally – prediction model induction and extraction of classification rules

Frequent Subgraph Extraction Example

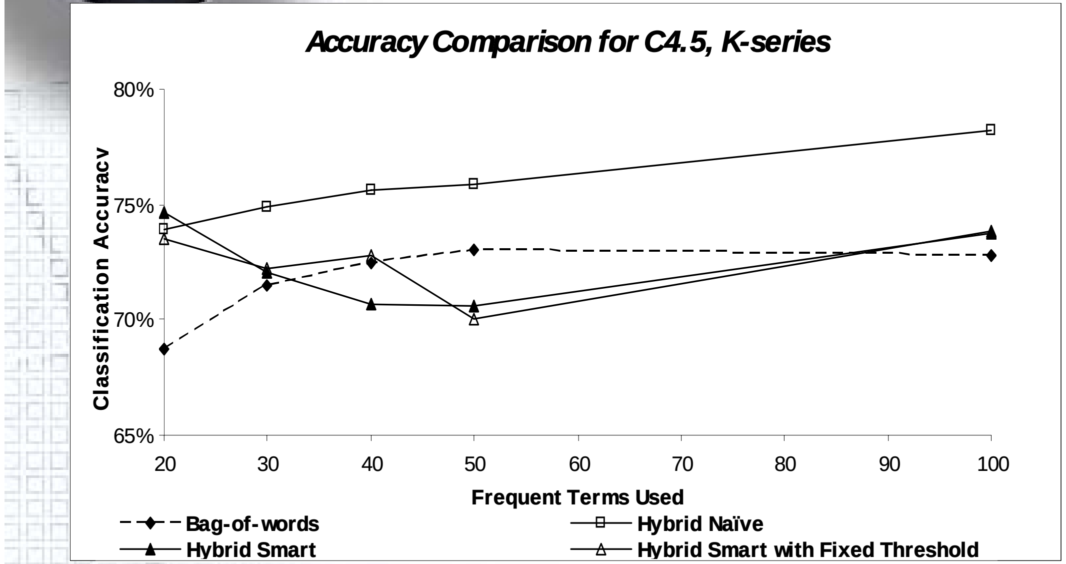
(based on the FSG algorithm by Kuramochi and Karypis, 2004)

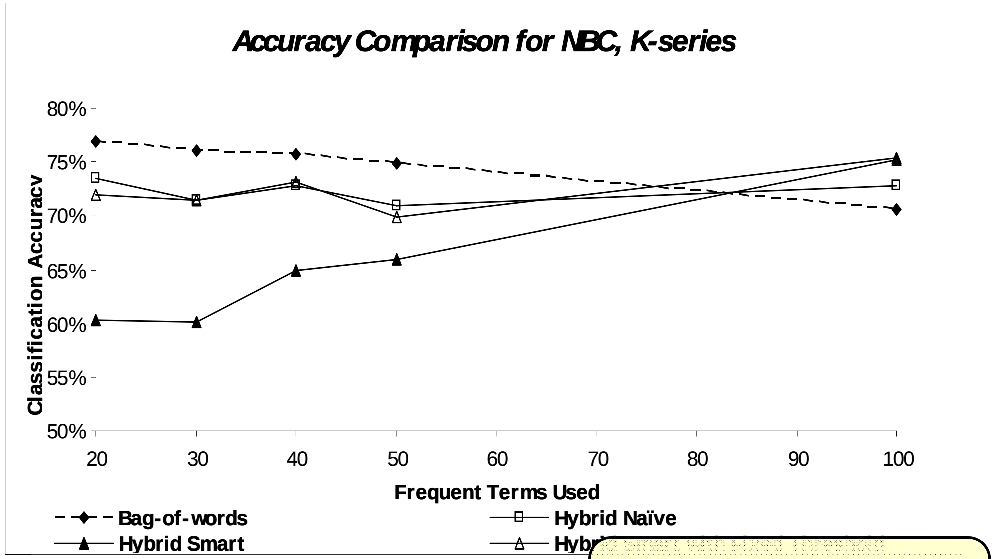


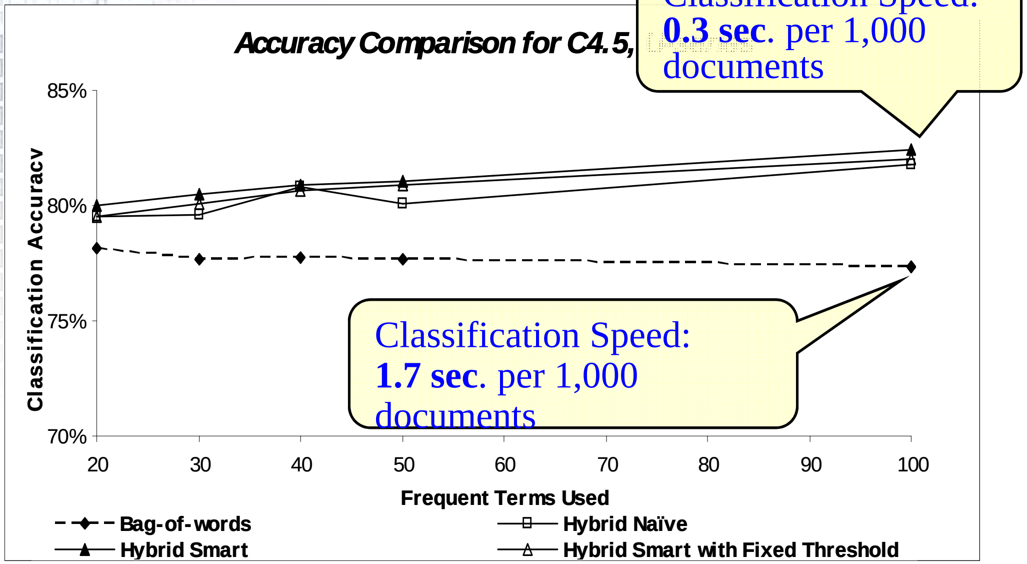
Comparative Evaluation

- Benchmark Data Sets
  - K-series (Source: Boley *et al*., 1999)
    - 2,340 documents and 20 categories
    - Documents in that collection 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
    - Known as "WebKB Dataset"
- 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

Categorization Accuracy and Speed

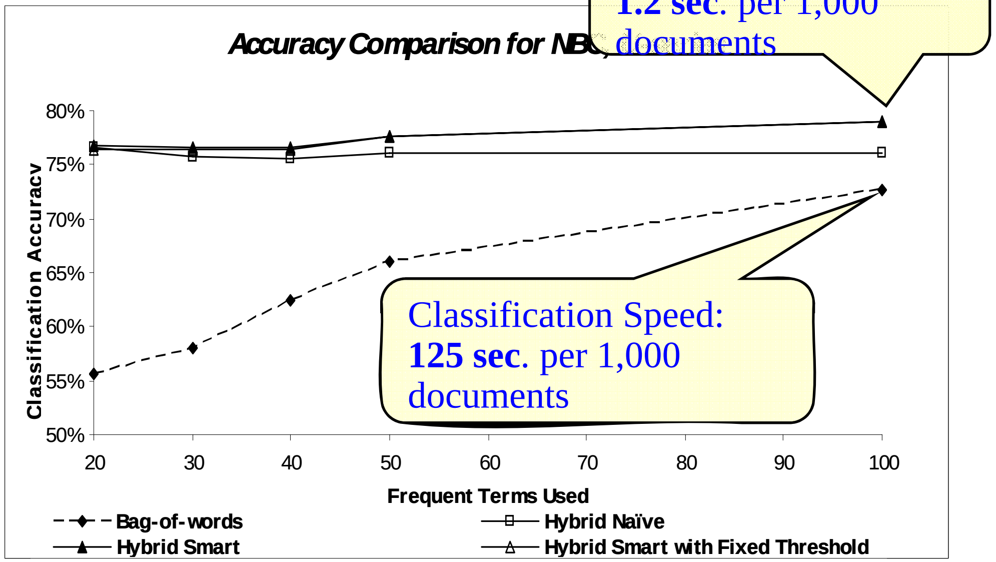






Classification Speed: **0.3 sec**. per 1,000 documents

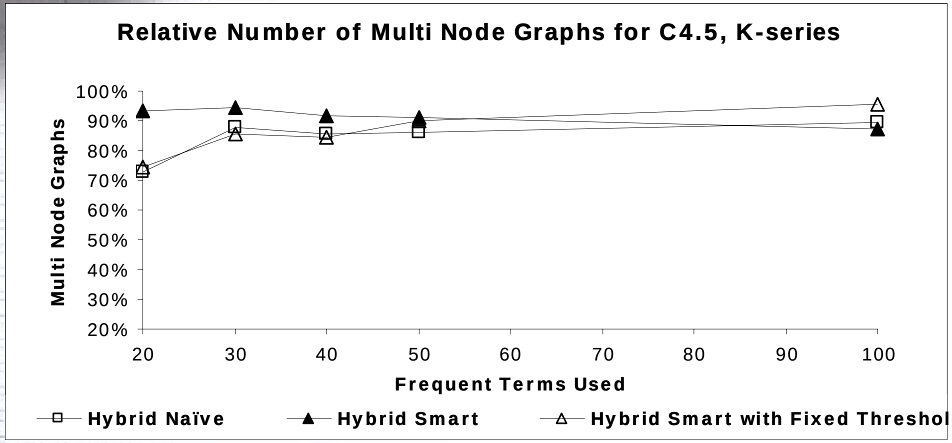
Classification Speed: **1.7 sec**. per 1,000 documents

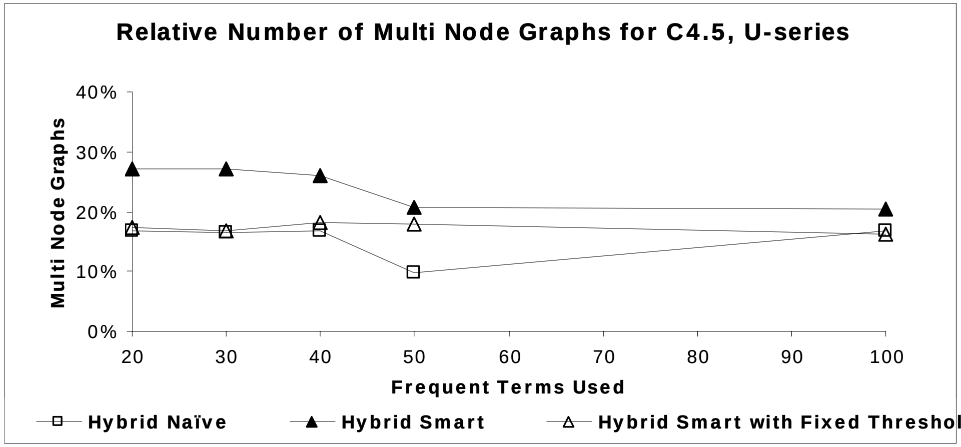


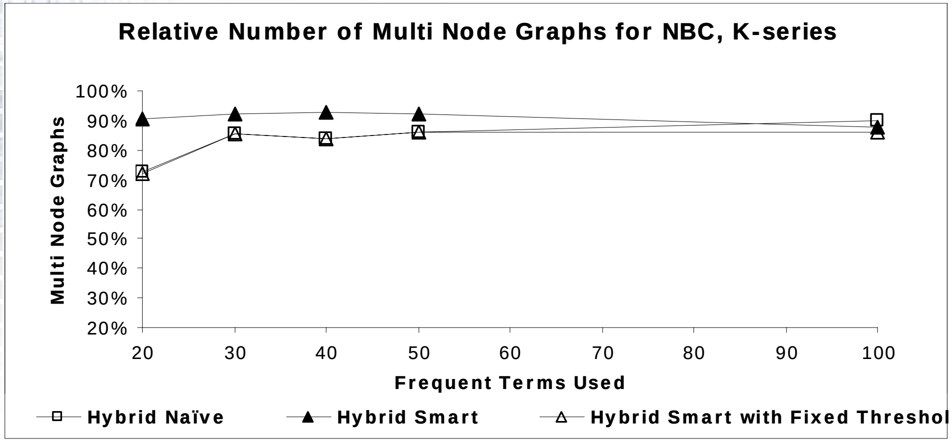
Classification Speed: **1.2 sec**. per 1,000 documents

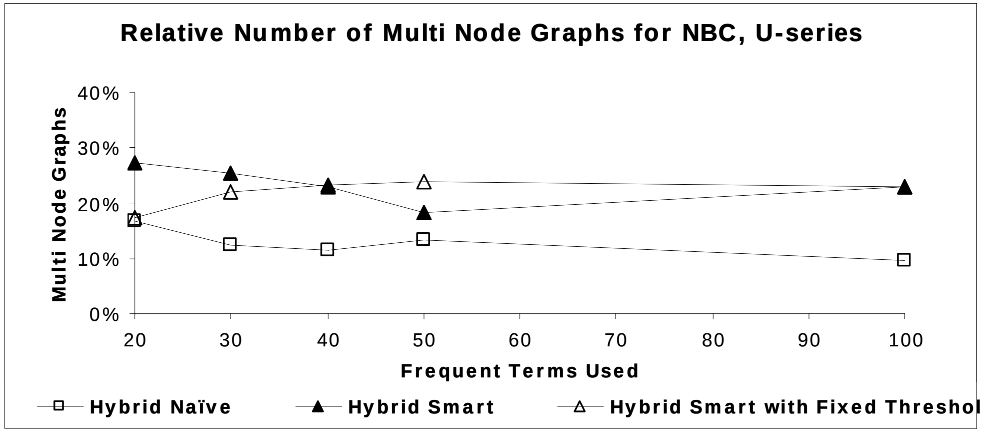
Classification Speed: **125 sec**. per 1,000 documents

Percentage of Multi-node Subgraphs









Litvak and Last (2008)

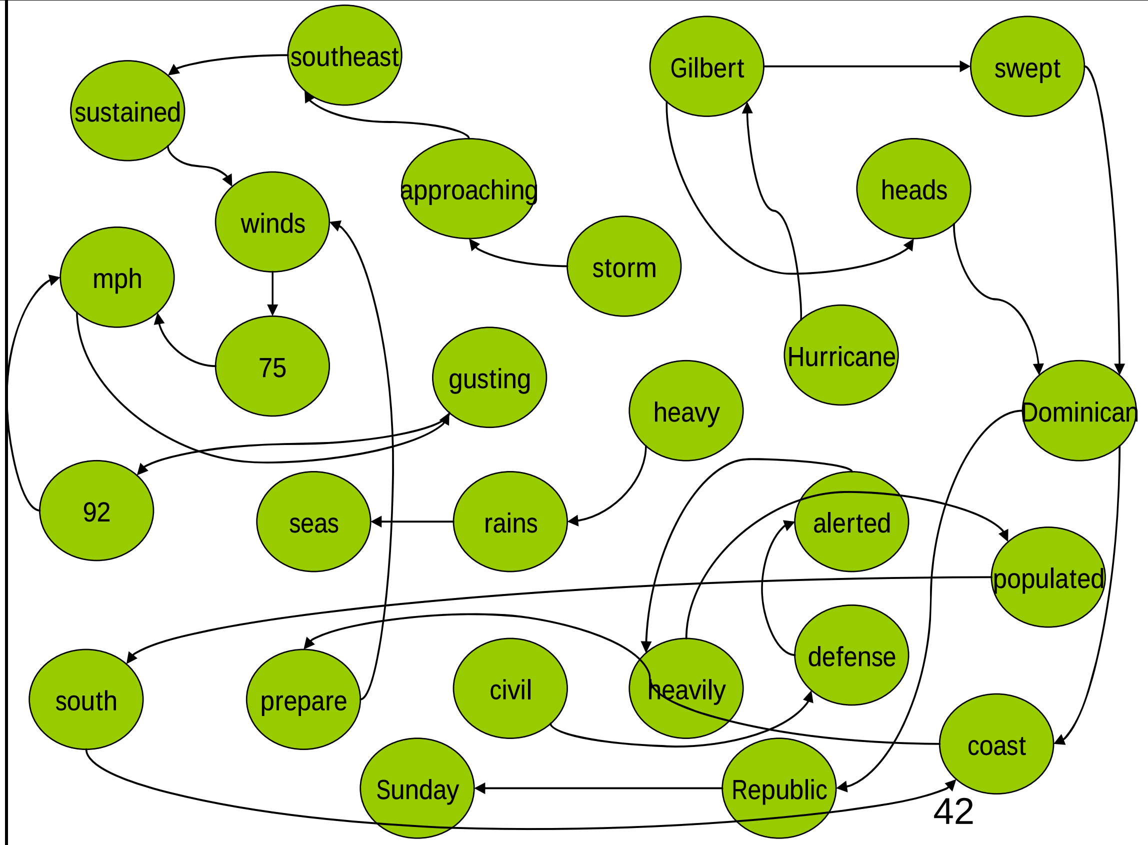
GRAPH-BASED KEYWORD EXTRACTION

Our methodology

- The *keyword* - is a word presenting in the document summary.
- *Document representation* - the ”simple” directed graph:
  - – Unique nodes – non-stop words
  - – Unlabeled edges - order-relationship
    - A → B ⇔ B appears after A in the same sentence
- *Keyword extraction* as a first stage of extractive summarization
  - – The most salient words (”keywords”) are extracted in order to generate a summary.

The “simple” graph-based document representation

Example:

Text	Graph
<title> Hurricane Gilbert Heads Toward Dominican Coast *</title>* *<TEXT>* Hurricane Gilbert swept toward the Dominican Republic Sunday, and the Civil Defense alerted its heavily populated south coast to prepare for high winds, heavy rains and high seas. The storm was approaching from the southeast with sustained winds of 75 mph gusting to 92 mph. *</TEXT>*	 southeast Gilbert swept sustained approaching heads winds mph storm 75 gusting Hurricane heavy Dominican 92 seas rains alerted populated south prepare civil heavily defense coast Sunday Republic

Keyword extraction
The supervised approach

- Training a *classification algorithm* on a repository of summarized documents.

- Each *node* in a document graph belongs to one of *two classes*:
  - YES - the word is included in the document extractive summary
  - NO - otherwise.

The Supervised approach (cont.)

The *features* used for nodes classification:

- *In Degree* – number of incoming edges
- *Out Degree* – number of outgoing edges
- *Degree* – total number of edges
- *Frequency* – term frequency of the word represented by node
- *Frequent words distribution* – $\in$ {0, 1}, equals to 1 iff Frequency $\geq$ threshold (*0.05*)
- *Location Score* – an average of location scores between all sentences (S(N)) containing the word N represented by node, where sentence location score is an reciprocal of the sentence location in text *(1/i)*
- *Tfidf Score* – the tf-idf score of the word represented by node. We used formula: $\frac{tf}{tf+1}\log_2\frac{|D|}{df}$
- *Headline Score* – $\in$ {0, 1}, equals to 1 iff document headline contains word represented by node

Feature extraction

Example:

- Node "Dominican":

•*In Degree* = 2

•*Out Degree* = 2

•*Degree* = 4

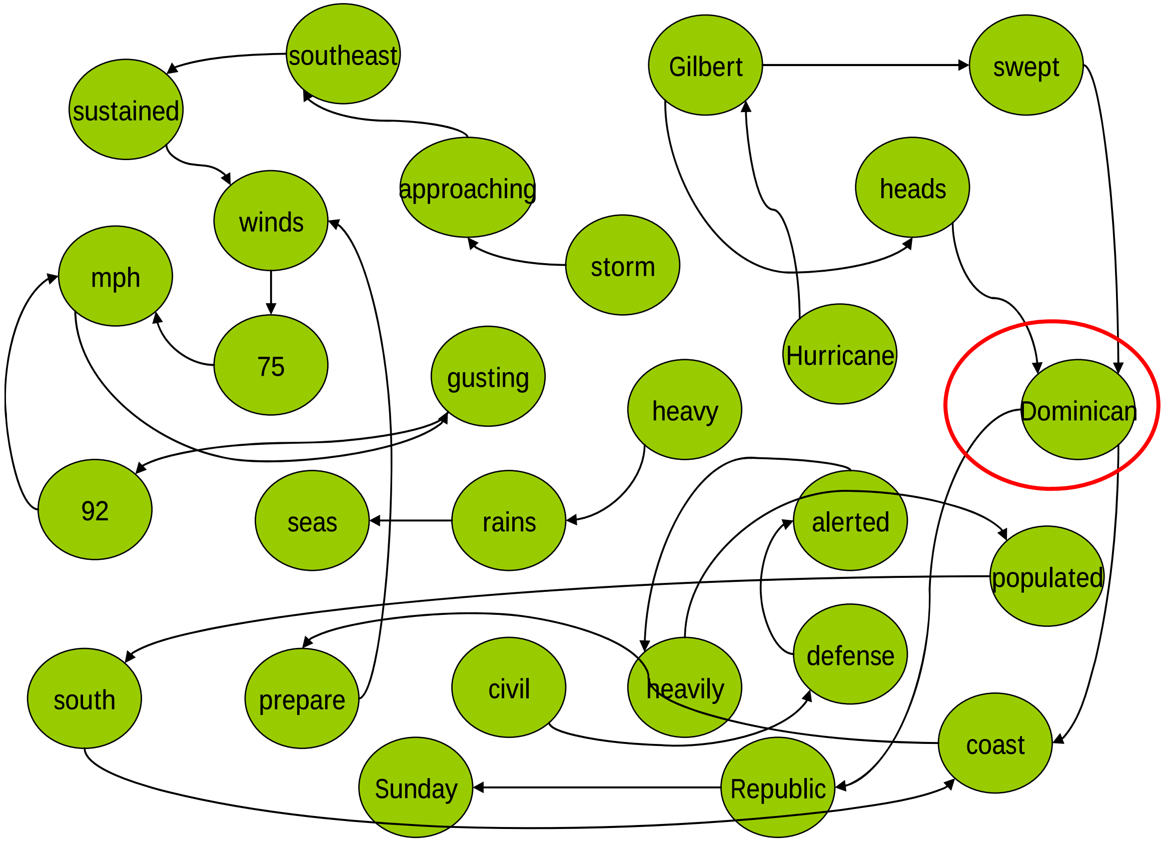
•*Frequency* = 2/27 = 0.074

•*Frequent words distribution* = 1

•*Location Score* = (1/1+1/2)/2 = 0.75

•*Tfidf Score* = (0.07/1.07)*$\log_2$(566/2) = 0.53

•*Headline Score* = 1



The unsupervised approach

- Unsupervised text unit extraction in the context of the text summarization task.
- No collection of summarized documents is needed
- We apply the HITS algorithm to document graphs.

HITS

Kleinberg, J.M. 1999.

- For each node, HITS produces two sets of scores - an ”authority” and a ”hub”:

$$HITS_A(V_i) = \sum_{V_j \in In(V_i)} HITS_H(V_j) \qquad (1)$$

$$HITS_H(V_i) = \sum_{V_j \in Out(V_i)} HITS_A(V_j) \qquad (2)$$

- For the total rank (H) calculation we used the following four functions:

$$H(V_i) = HITS_A(V_i)$$

$$H(V_i) = HITS_H(V_i)$$

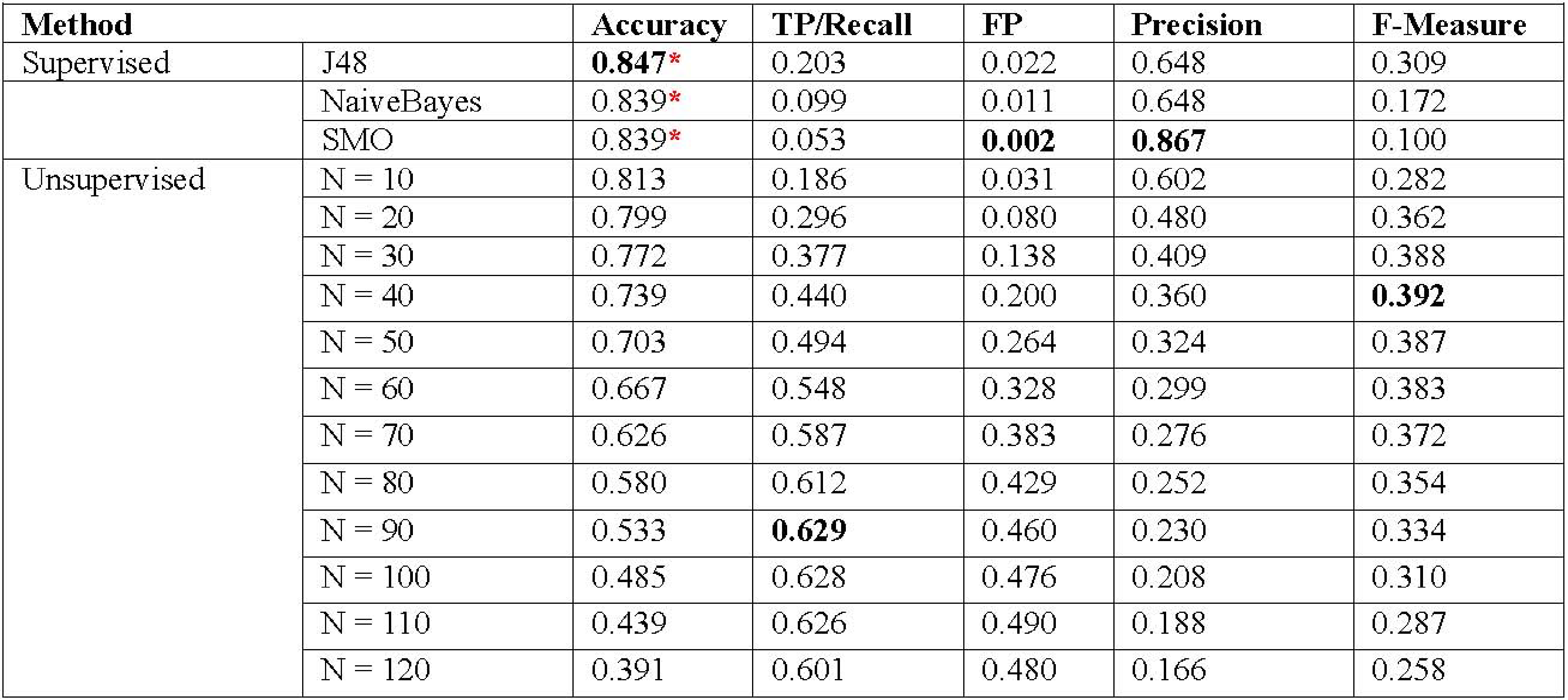
$$H(V_i) = avg\{HITS_A(V_i), HITS_H(V_i)\}$$

$$H(V_i) = \max\{HITS_A(V_i), HITS_H(V_i)\}$$

Experimental results

- DUC, 2002 collection:
  - 566 English texts along with 2-3 summaries per document on average.
  - The size (|V|) of syntactic graphs extracted from these texts is 196 on average, varying from 62 to 876.

Comparison of supervised and unsupervised approaches



Method		Accuracy	TP/Recall	FP	Precision	F-Measure
Supervised	J48	**0.847***	0.203	0.022	0.648	0.309
	NaiveBayes	0.839*	0.099	0.011	0.648	0.172
	SMO	0.839*	0.053	**0.002**	**0.867**	0.100
Unsupervised	N = 10	0.813	0.186	0.031	0.602	0.282
	N = 20	0.799	0.296	0.080	0.480	0.362
	N = 30	0.772	0.377	0.138	0.409	0.388
	N = 40	0.739	0.440	0.200	0.360	**0.392**
	N = 50	0.703	0.494	0.264	0.324	0.387
	N = 60	0.667	0.548	0.328	0.299	0.383
	N = 70	0.626	0.587	0.383	0.276	0.372
	N = 80	0.580	0.612	0.429	0.252	0.354
	N = 90	0.533	**0.629**	0.460	0.230	0.334
	N = 100	0.485	0.628	0.476	0.208	0.310
	N = 110	0.439	0.626	0.490	0.188	0.287
	N = 120	0.391	0.601	0.480	0.166	0.258

- We consider unsupervised model based on extracting top *N* ranked words for different values of $10 \leq N \leq 120$.
- Set from *top 2* features: *Frequent words distribution* and *In Degree* is used for NBC

SUMMARY

Selected Publications

- A. Schenker, M. Last, H. Bunke, A. Kandel, "Classification of Web Documents Using Graph Matching", *International Journal of Pattern Recognition and Artificial Intelligence*, Special Issue on Graph Matching in Computer Vision and Pattern Recognition, Vol. 18, No. 3, 2004, pp. 475-496.
- A. Schenker, H. Bunke, M. Last, A. Kandel, "Graph-Theoretic Techniques for Web Content Mining", *World Scientific*, 2005.
- A. Markov, M. Last, "A Simple, Structure-Sensitive Approach for Web Document Classification", *Atlantic Web Intelligence Conference (AWIC2005)*, Lodz, Poland, June 2005.
- A. Markov, M. Last, and A. Kandel, “Fast Categorization of Web Documents Represented by Graphs”, in Advances in Web Mining and Web Usage Analysis, O. Nasraoui, *et al.* (Eds), *Springer Lecture Notes in Computer Science (LNCS/LNAI)*, Vol. 4811, 2007, pp. 56-71.
- A. Markov, M. Last, and A. Kandel, “The Hybrid Representation Model for Web Document Classification”, *International Journal of Intelligent Systems*, Vol. 23, No. 6, pp. 654-679, 2008.
- M. Litvak and M. Last, "Graph-Based Keyword Extraction for Single-Document Summarization", *Proceedings of the 2nd Workshop on Multi-source, Multilingual Information Extraction and Summarization (MMIES2)*, Manchester, UK, August 23, 2008, pp. 17–24.

Future Research

- Enhancing graph representations of text and web documents
  - Utilizing POS tagging
  - Concept fusion based on available ontologies
  - Implementing graph representations for more languages
- Identification of the most relevant sections in long documents, online forums, etc.
- Cross-lingual summarization of text documents
- Topic detection and tracking in the web content
- Opinion and sentiment mining

Thank you!