אוניברסיטת בן-גוריון בנגב
Ben-Gurion University of the Negev

NIACI
2001
National Institute for Applied Computational Intelligence

Model-Based Classification of Web Documents Represented by Graphs

Alex Markov and Mark Last

Department of Information Systems Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel

Abraham Kandel

National Institute for Applied Computational Intelligence

University of South Florida, Tampa, FL, USA

E-mail: mlast@bgu.ac.il
Home Page: http://www.ise.bgu.ac.il/faculty/mlast/

WebKDD 2006 Workshop on Knowledge Discovery on the Web, Aug. 20, 2006, at KDD 2006, Philadelphia, PA, USA

Content

- Introduction and Motivation
- Graph-based Representation of Web Documents
- The Hybrid Methodology for Web Document Representation and Classification
  - The Naïve Approach
  - The Smart Approach
  - The Smart Approach with Fixed Threshold
- Comparative Evaluation
- Conclusions and Future Research

Motivation

- Most of Web document classification algorithms
  - Treat web documents the same way as text documents
    - HTML tags are completely ignored
- The popular Vector-Space model
  - Ignores the word position in the document
  - Ignores the order of words in the document
- Solution – structure-sensitive document representation
  - Graph representation in this research

Text Categorization (TC)

Relevant Definitions

- 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 domain of documents and $C = (c_1, \ldots, c_{|C|})$ is set of pre-defined categories (classes)

- *Single Label TC* – only one category can be assigned to each document
- *Multi Label TC* – overlapping categories allowed
- *Ranking* categorization
  - Degree of relevance of every document to each category is calculated

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

title

```
<!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
```

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>
```

text

Graph Based Document Representation - Preprocessing

TITLE

CNN International

Stop word removal

Stemming

Text

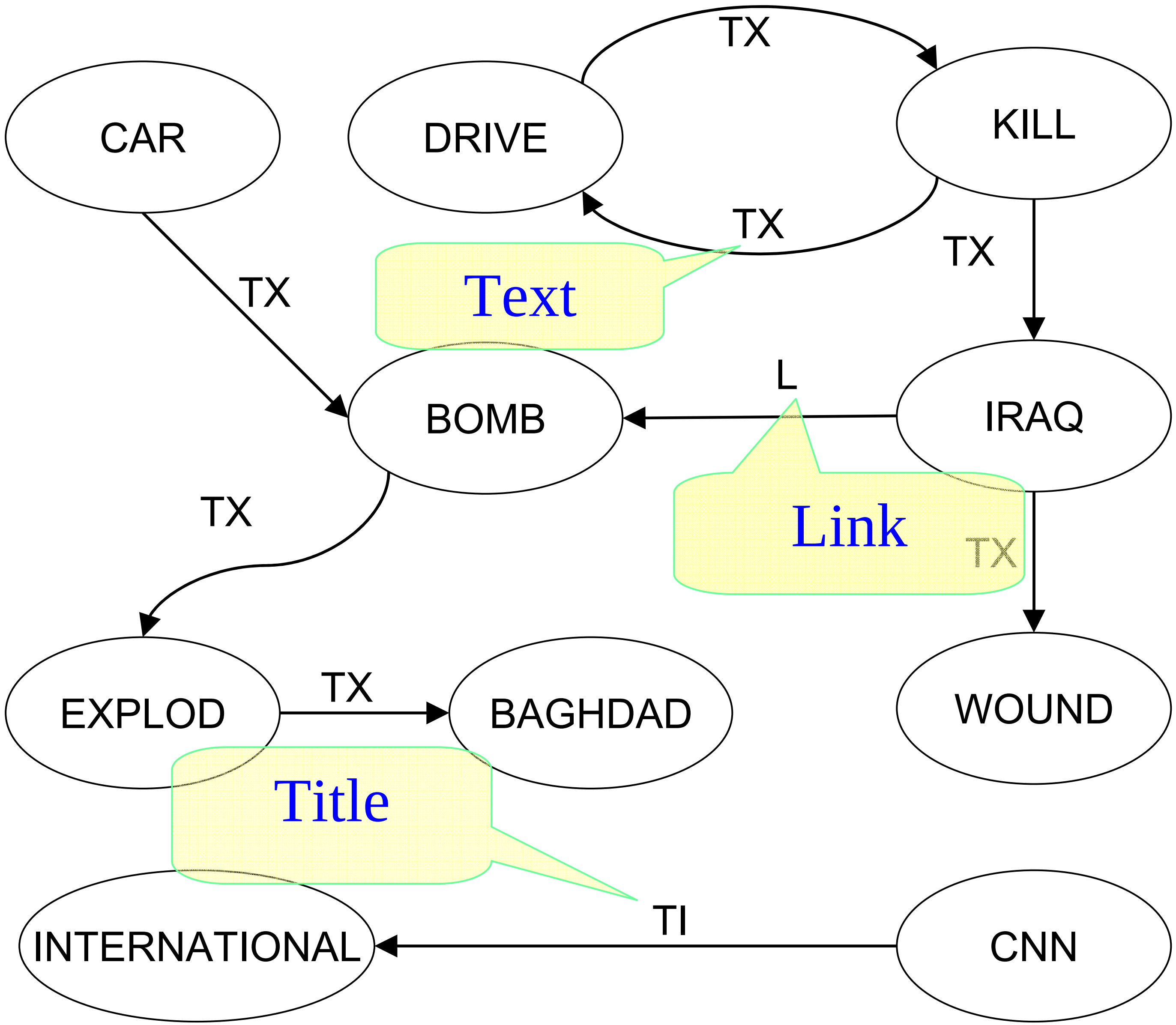
car bomb explod Baghdad restaurant, kill Iraq wound , police official . aide office Iraq Prime Minister Ibrahim al-Jaafari drive kill drive shooting.

Links

Iraq bomb: dead, wound .
FULL STORY.

Graph Based Document Representation – Graph Construction

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



Web Document Classification with Graph-Based Models

- Advantages (Schenker *et al*., 2004)
  - Keep HTML structure information
  - Retain original order of words
- Limitation
  - Can work only with "lazy" classifiers, which have a very low classification speed
    - Example: k-Nearest Neighbors classifier
- Conclusion
  - Graph models cannot be used directly for <u>model-based</u> classification of web documents
- Solution
  - The **hybrid approach**: represent a document as a <u>vector of sub-graphs</u>

Graph Based Document Representation – Subgraphs Extraction

- ***Naïve Method***
  - Input:
    - ***G*** - Training set of directed, unique nodes graphs
    - t_{min} – Threshold (minimum sub-graph frequency)
  - Output:
    - Set of classification-relevant sub-graphs
  - Process:
    - For each class find frequent sub-graphs $\boldsymbol{SCF} > \boldsymbol{t}_{min}$
    - Combine all sub-graphs into one set

Subgraph Class Frequency

- **Classification-Relevant Sub-Graphs** are frequent in a specific category

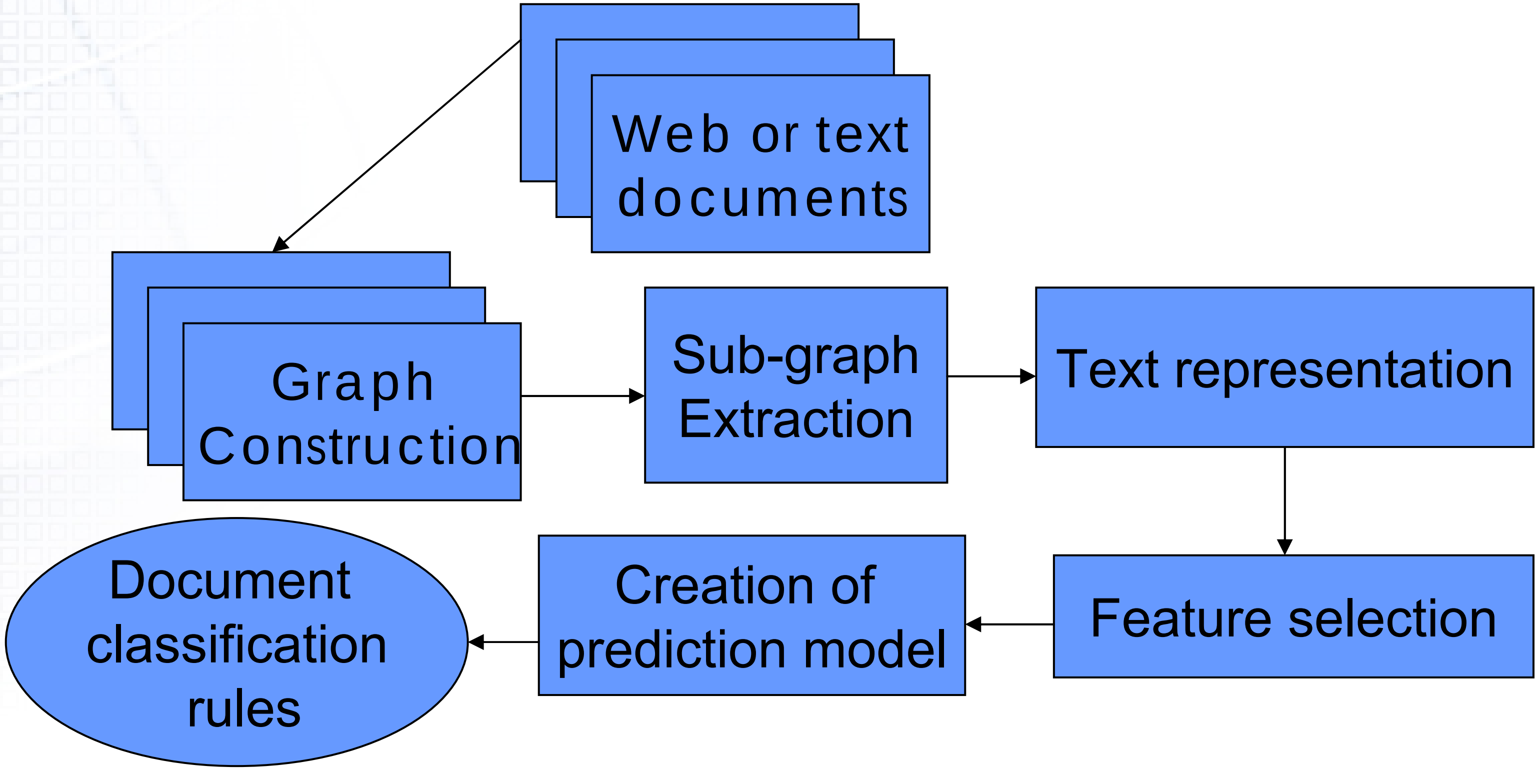
Graph Based Document Representation – Subgraphs Extraction

- ***Smart Method***
  - 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 sub-graphs $\boldsymbol{CR} > \boldsymbol{CR}_{min}$
    - Combine all sub-graphs into one set
- **Classification-Relevant Sub-Graphs** are more frequent in a specific category than in other categories

Graph Based Document Representation – Subgraphs Extraction

- ***Smart with Fixed Threshold Method***
  - Input
    - ***G*** – training set of directed, unique nodes graphs
    - t_{min} – Threshold (minimum sub-graph frequency)
    - CR_{min} - Minimum Classification Rate
  - Output
    - Set of classification-relevant sub-graphs
  - Process:
    - For each class find sub-graphs $\boldsymbol{SCF > t_{min}}$ and $\boldsymbol{CR > CR_{min}}$
    - Combine all sub-graphs into one set
- **Classification-Relevant Sub-Graphs** are frequent in a specific category *and* not frequent in other categories

Predictive Model Induction with Hybrid Representation



Frequent Subgraphs 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: Algorithm

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

$F^0 \leftarrow$ Detect all frequent 1 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 Subgraphs Extraction: Complexity

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

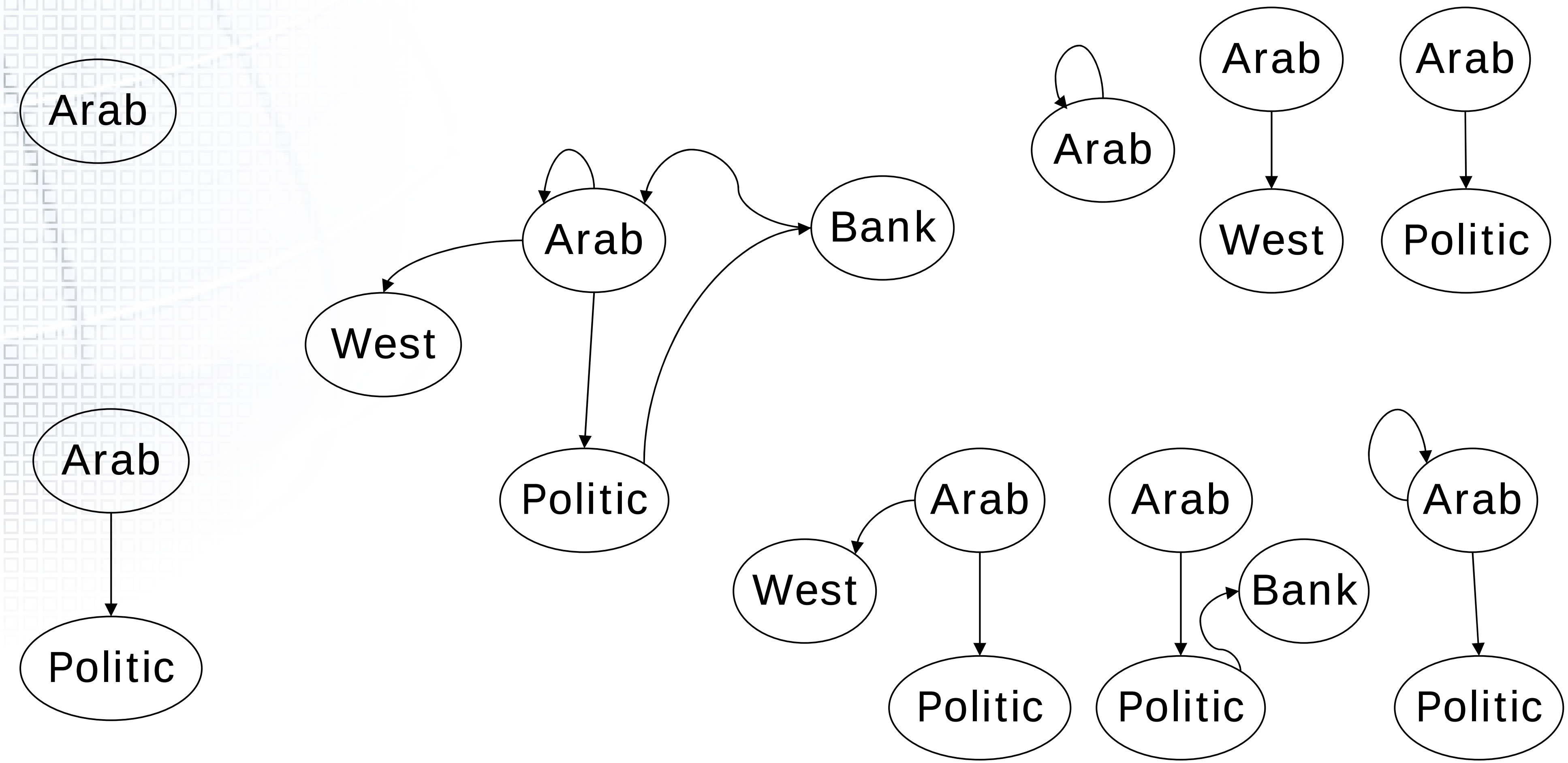
Isomorphism between graphs $G_1=(V_1,E_1,\alpha_1,\beta_1)$ and $G_2=(V_2,E_2,\alpha_2,\beta_2)$ can be found by two simple actions:

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

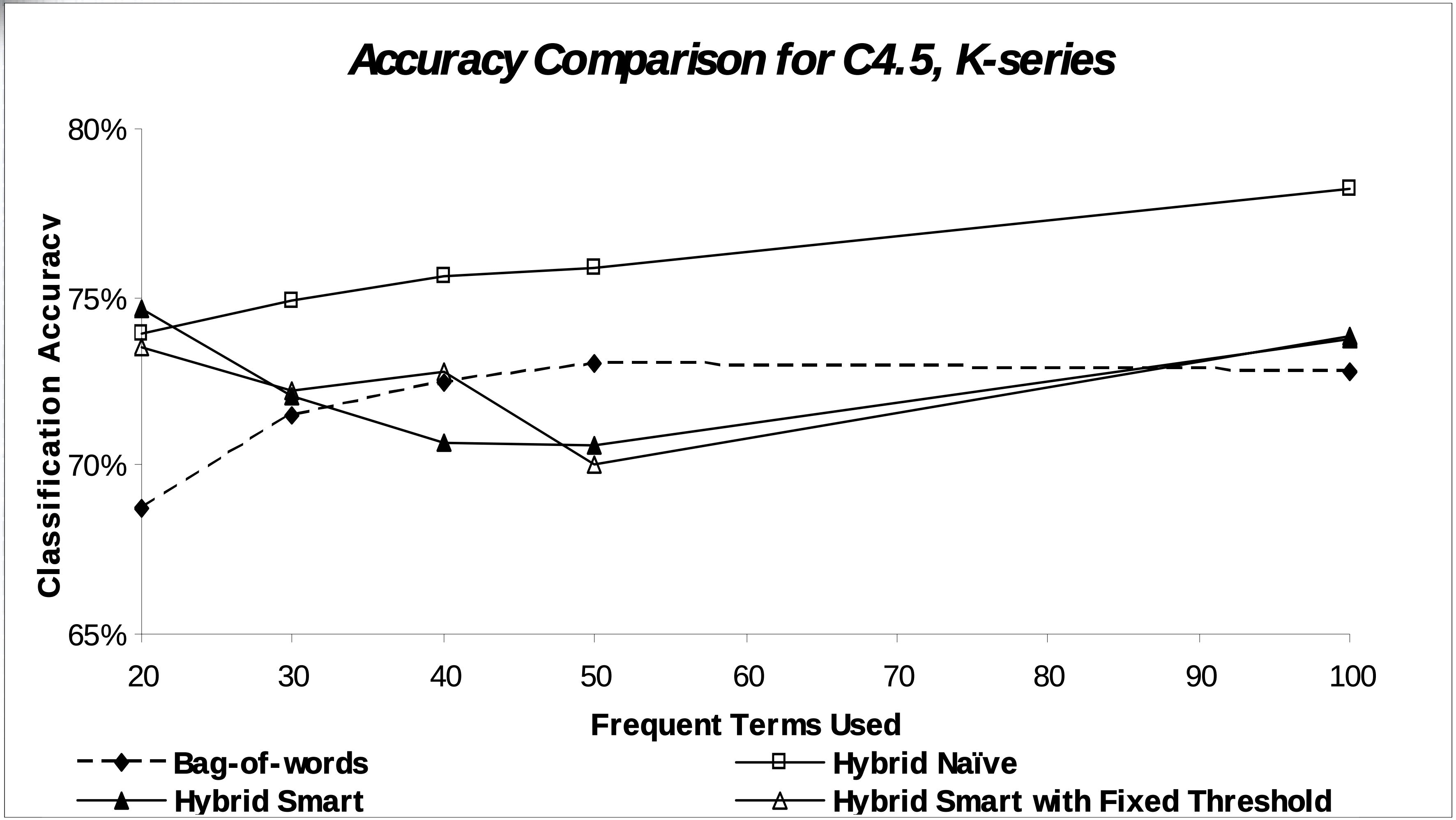
Subgraphs **Document Graph** **Extensions**



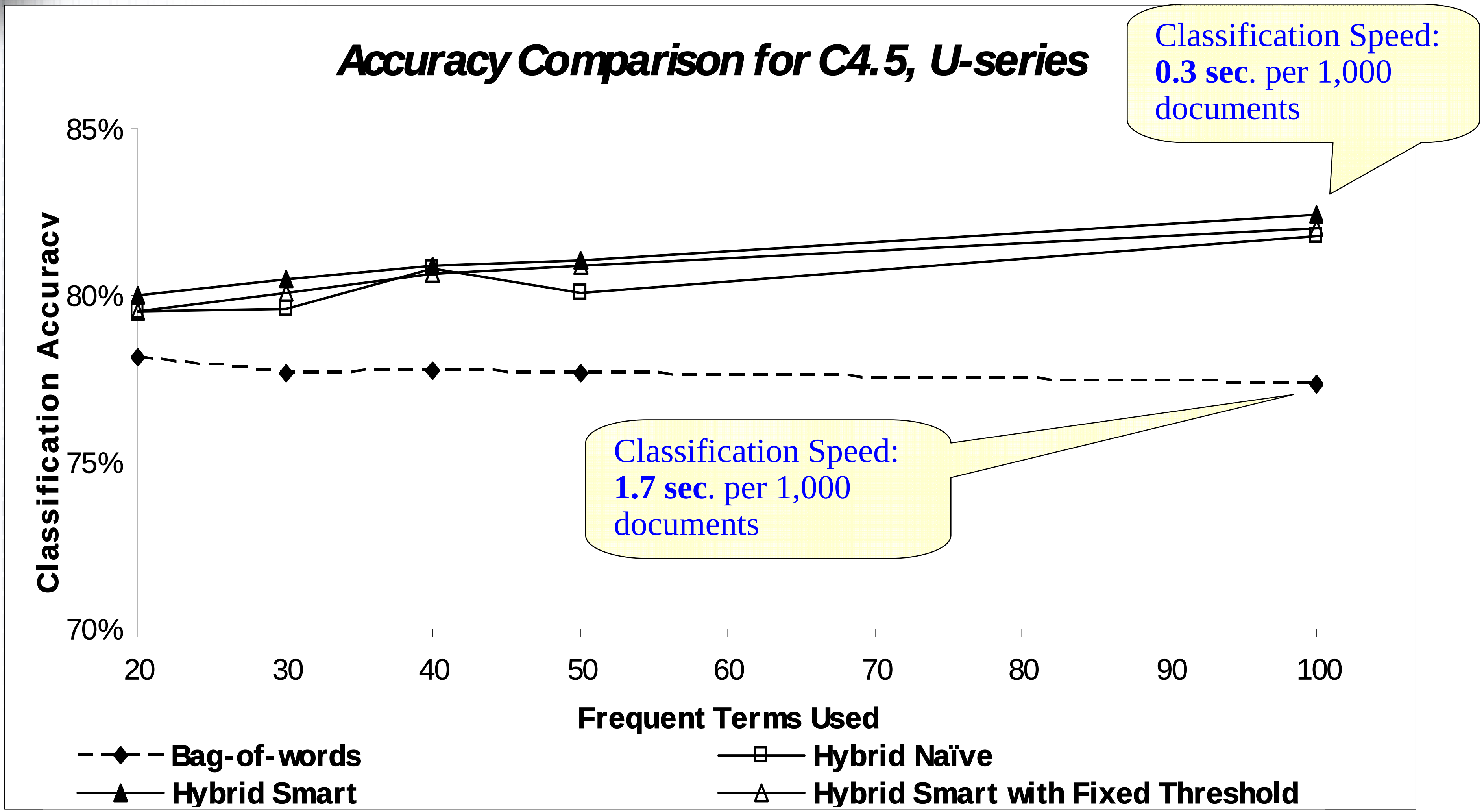
Comparative Evaluation

- Benchmark Data Sets
  - K-series
    - 2,340 documents and 20 categories
    - Documents in those collections were originally news pages hosted at Yahoo
  - U-series
    - 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

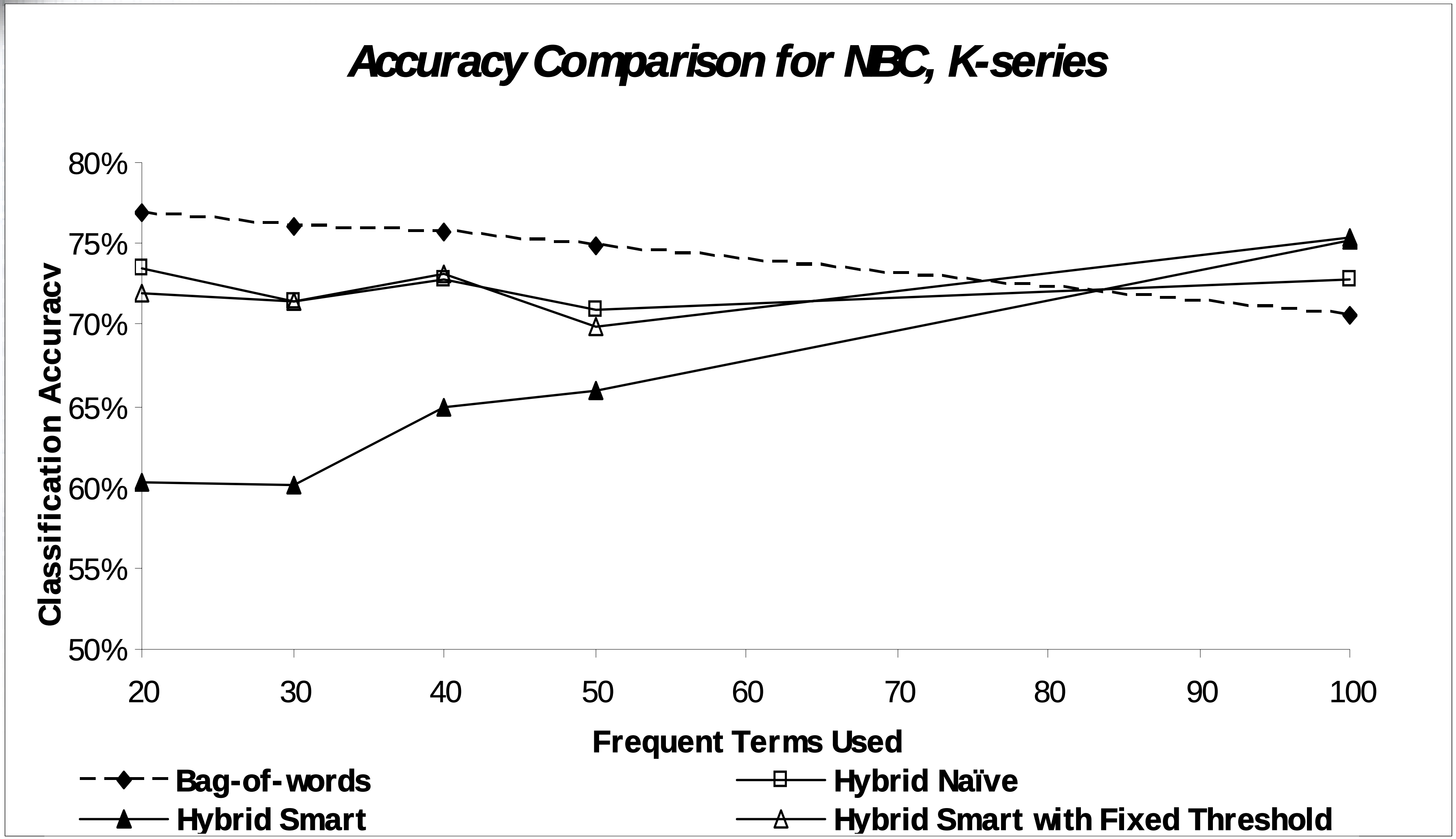
Classification Results with C4.5– K series data set



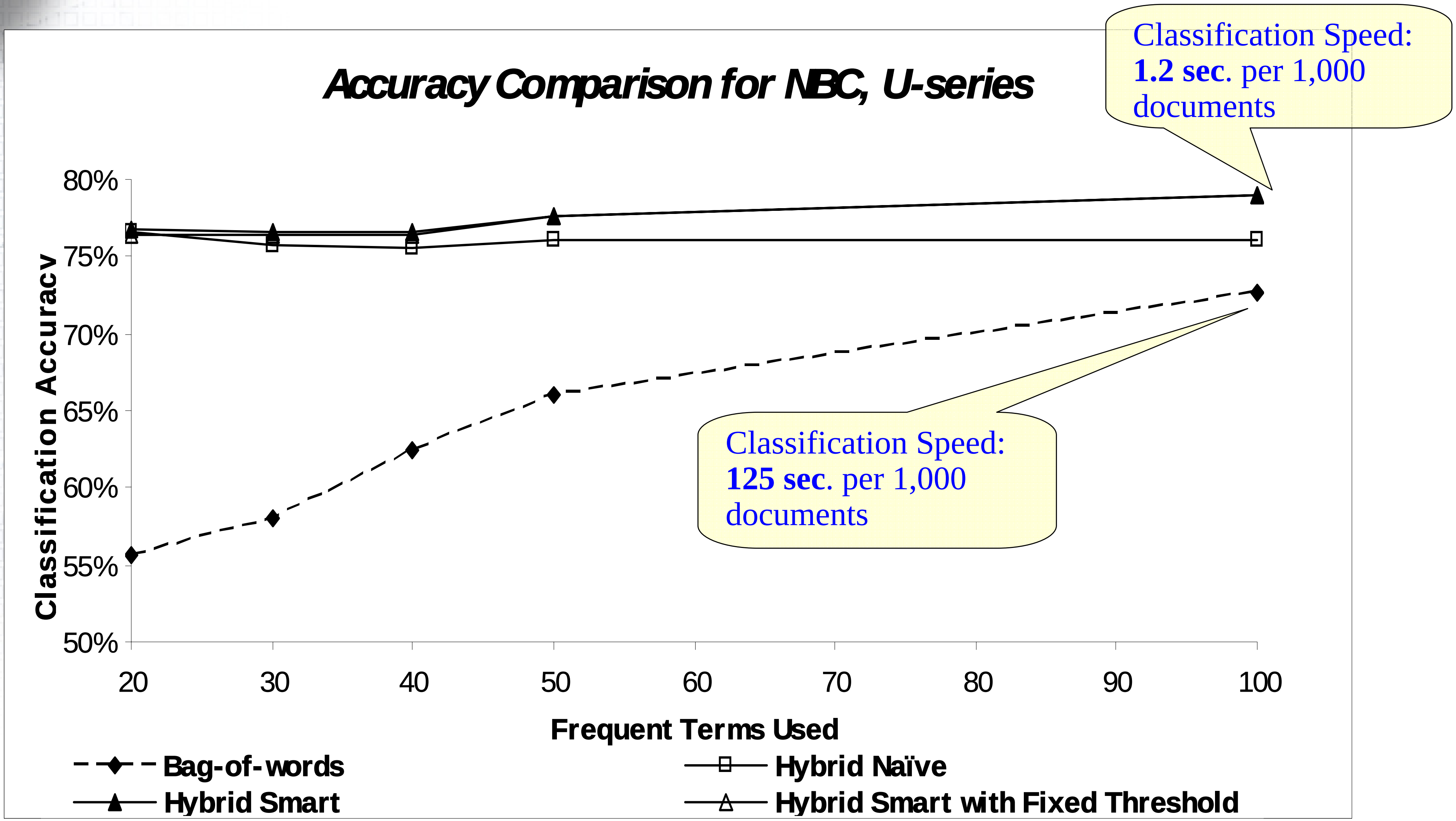
Classification Results with C4.5– U series data set



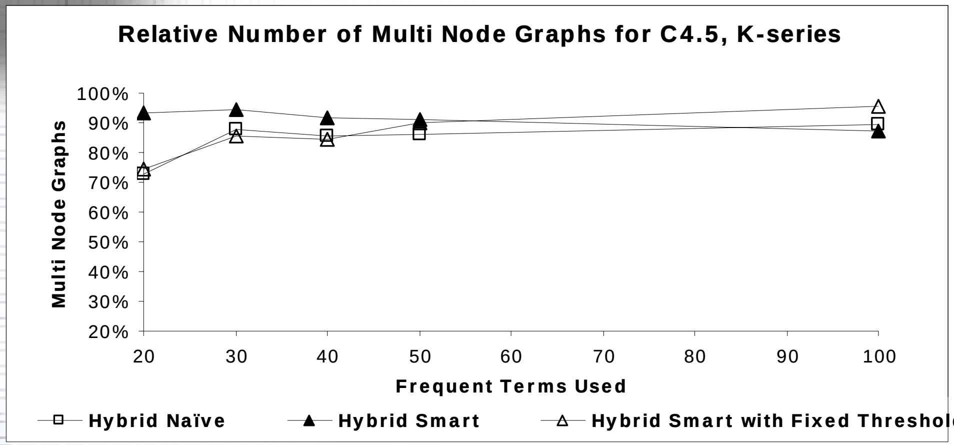
Classification Results with Naïve Bayes – K series data set

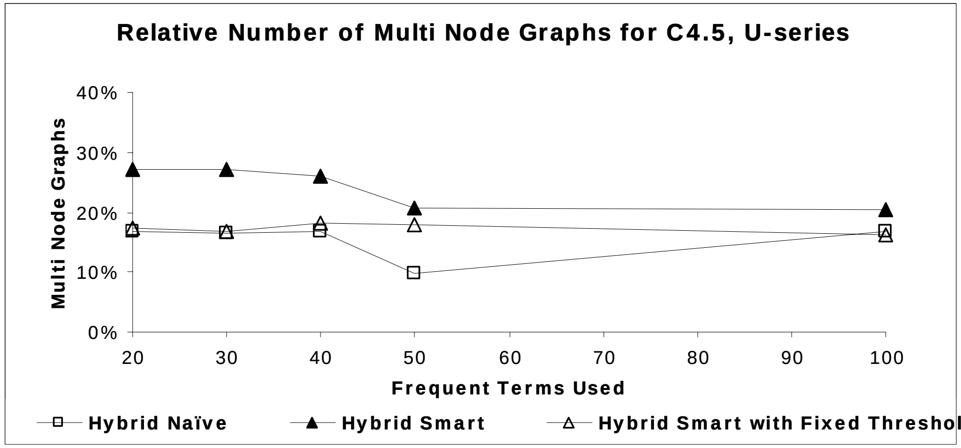


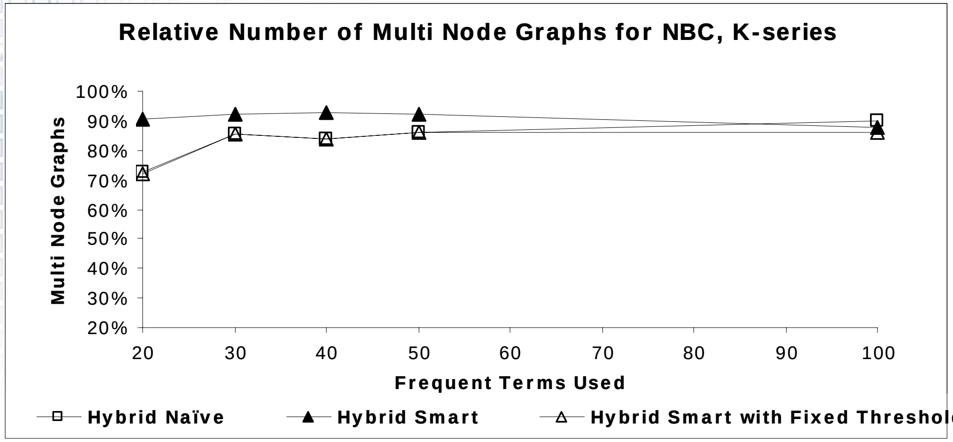
Classification Results with Naïve Bayes – U series data set

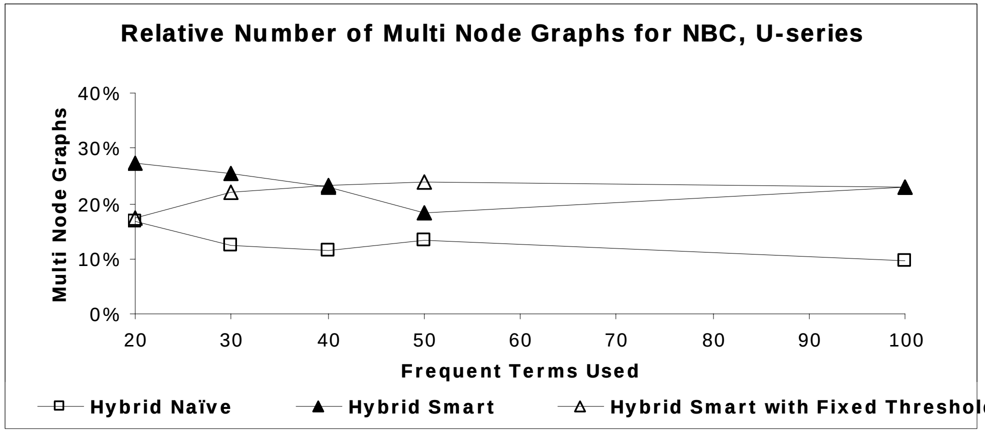


Percentage of Multi-node Subgraphs









Summary

- Different document representations were empirically compared in terms of classification accuracy and execution time
- The proposed hybrid methods were found to be more accurate in most cases and generally much faster than their vector-space and graph-based counterparts

Future research

- Finding optimal parameters for sub-graph extraction:
  - Graph size *N*
  - t_{min} for Naïve extraction
  - CR_{min} for Smart extraction
- Applying the hybrid methodology to additional classifiers
- Applying the hybrid methodology to unsupervised learning (clustering)

Thank you!

Selected References

- M. Kuramochi and G. Karypis, "An Efficient Algorithm for Discovering Frequent Subgraphs", IEEE Transactions on Knowledge and Data Engineering, Volume 16 , Issue 9, September 2004.
- 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.
- 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 and M. Last, "Efficient Graph-Based Representation of Web Documents", Proceedings of the Third International Workshop on Mining Graphs, Trees and Sequences (MGTS2005), October 7, 2005, Porto, Portugal.
- 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), Singapore, April 9, 2006.