
File Classification Using Sub-Sequence Kernels

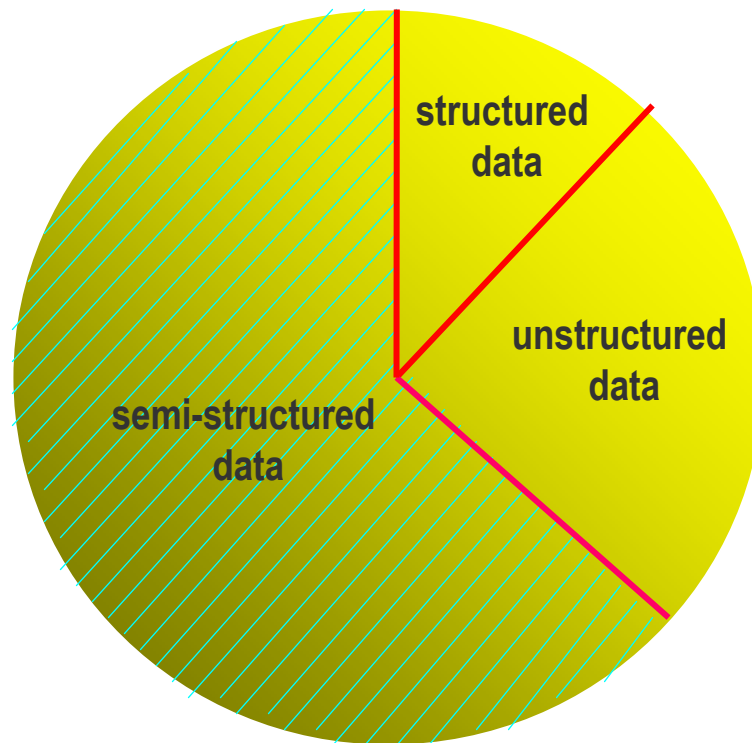
Olivier de Vel
Computer Forensics Group
DSTO, Australia

Presentation:

- ✓ **Document mining and computer forensics**
- ✓ **Semi-structured documents**
- ✓ **Broadband k -spectrum kernel**
- ✓ **Coloured generalized suffix tree representation**
- ✓ **Experimental methodology**
- ✓ **Results and Conclusion**

Computer Forensics: Focus on Document Mining

Eg, disk surface contents



Presentation:

- ✓ Document mining and computer forensics
- ✓ **Semi-structured documents**
- ✓ Broadband k -spectrum kernel
- ✓ Coloured generalized suffix tree representation
- ✓ Experimental methodology
- ✓ Results and Conclusion

Semi-Structured Documents

- **May or may not include natural language text**
- **May be compound documents**
- **Include low-level, short-range structures. For example:**
 - **byte block identifiers**
 - **mark-up tokens**

Semi-Structured Documents

⇒ Traditional text mining techniques generally not appropriate.

Presentation:

- ✓ Document mining and computer forensics
- ✓ Semi-structured documents
- ✓ **Broadband k -spectrum kernel**
- ✓ Coloured generalized suffix tree representation
- ✓ Experimental methodology
- ✓ Results and Conclusion


Broadband k -Spectrum Kernel:

Use an extension to the string kernel.

$x_i =$ BOOK OOPERPPPO...

String (k -Spectrum) Kernel:

- Is the set of all k -length contiguous subsequences that a sequence x_i (of alphabet T , size $|T|=l$) contains
- Feature space dimension is up to $|T|^k$ (eg, if $k=4$ and $|T| = 256 \Rightarrow$ number of dimensions $\approx 10^9$)



 $(k=4)$

 BOOK

 OOKO

 OKOO

 KOOP

 OOPE

 etc....

Vector representation: $(\varphi_{\kappa_1}(x_i), \varphi_{\kappa_2}(x_i), \dots, \varphi_{\kappa_l}(x_i))$
 where $\kappa_1, \kappa_2, \dots, \kappa_l \in T^k$

Broadband k -Spectrum Kernel:

Broadband k -Spectrum Kernel:

- Is the set of all $(0, 1, ..k)$ -length contiguous subsequences that a sequence (of alphabet T) contains
- Feature space dimension is even larger!

BOOKOOPERPPO...



$(k=4)$

B, BO, BOO, BOOK
O, OO, OOK, OOKO
O, OK, OKO, OKOO
K, KO, KOO, KOOP
O, OO, OOP, OOPE
etc....

Broadband k -Spectrum Kernel:

Matrix representation, $\varphi_{(BB)}(\mathbf{x}_i)$:

$$\begin{pmatrix} \varphi_{\kappa_1^{(1)}}(\mathbf{x}_i) & \varphi_{\kappa_2^{(1)}}(\mathbf{x}_i) & \dots & \varphi_{\kappa_l^{(1)}}(\mathbf{x}_i), \\ \varphi_{\kappa_1^{(2)}}(\mathbf{x}_i) & \varphi_{\kappa_2^{(2)}}(\mathbf{x}_i) & \dots & \varphi_{\kappa_l^{(2)}}(\mathbf{x}_i), \\ \vdots & \vdots & \vdots & \vdots \\ \varphi_{\kappa_1^{(2)}}(\mathbf{x}_i) & \varphi_{\kappa_2^{(2)}}(\mathbf{x}_i) & \dots & \varphi_{\kappa_l^{(2)}}(\mathbf{x}_i) \end{pmatrix}$$

where $\kappa_i^{(j)} \in T^j$ for $i = 1, 2, \dots, l$

Broadband k -Spectrum Kernel: What's the Big Deal?

This allows us to define the inner product between two input sequences, x_r and x_s as:

$$\text{GramMatrix}(x_r, x_s) \equiv \Phi_{(BB)}(x_r) \cdot \Phi_{(BB)}(x_s)$$

and use the “kernel trick” when maximising the margin of separation between two classes. The decision function for the SVM learning algorithm is:

$$f(x) = \text{sign}\left(\sum \alpha_i y_i [\Phi_{(BB)}(x_r) \cdot \Phi_{(BB)}(x_s)] + b\right)$$

in the transformed feature space rather than the input feature space.

Presentation:

- ✓ Document mining and computer forensics
- ✓ Semi-structured documents
- ✓ Broadband k -spectrum kernel
- ✓ Coloured generalized suffix tree representation
- ✓ Experimental methodology
- ✓ Results and Conclusion

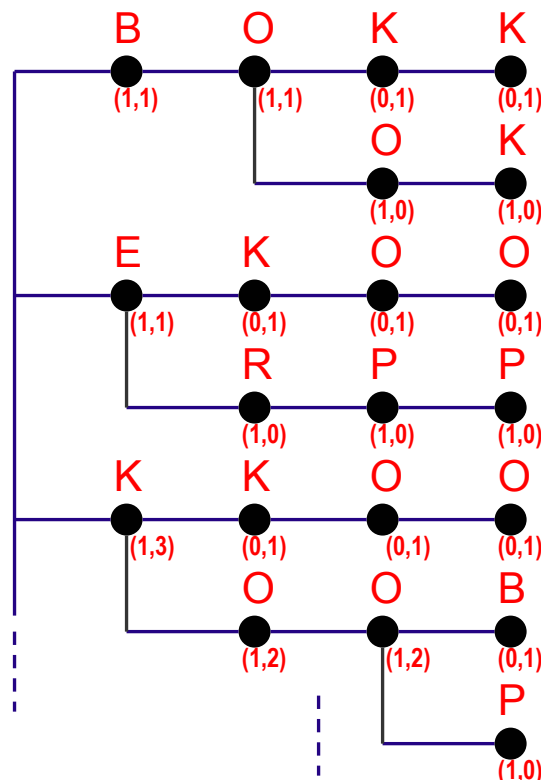
Coloured Generalized Suffix Tree (CGST):

The CGST is a labelling of the GST representation of a sequence of symbols:

It stores a count vector to be stored at each node in the GST – each vector element value equals the frequency of the subsequence in the sequence.

BOOKOOPERPPPO PPEKOOBOKKOO

(k=4)



CGST: Computing the Kernel (Gram) Matrix

For N input sequences, the $N \times N$ Gram matrix G is computed as follows:

- initialize $G_{r,s} = 0 \quad \forall x_r \text{ and } x_s$
- traverse the CGST
- at each node, compute the product of each vector element pair and sum to $G_{r,s}$

Presentation:

- ✓ Document mining and computer forensics
- ✓ Semi-structured documents
- ✓ Broadband k-spectrum kernel
- ✓ Coloured generalized suffix tree representation
- ✓ **Experimental methodology**
- ✓ Results and Conclusion

Experimental Methodology: Corpus

Document corpus for a controlled experiment :

- **5 document classes (MSWord, JPEG, MSEXcel, Java source, Adobe PDF)**
- **465 documents (0.11KB min. to 523.3KB max. size)**
- **data are scaled to unit standard deviation and zero mean**

Experimental Methodology: Classifier

- **SVM^{light}** as the (two-way) classifier,
- Obtain two-way categorisation matrix for each document type category, using 10-fold cross-validation sampling,
- Calculate per-document-type category classification performance statistics – precision, recall and F_1 statistics.

Presentation:

- ✓ Document mining and computer forensics
- ✓ Semi-structured documents
- ✓ Broadband k-spectrum kernel
- ✓ Coloured generalized suffix tree representation
- ✓ Experimental methodology
- ✓ **Results and Conclusion**

Experiments: Performance Results

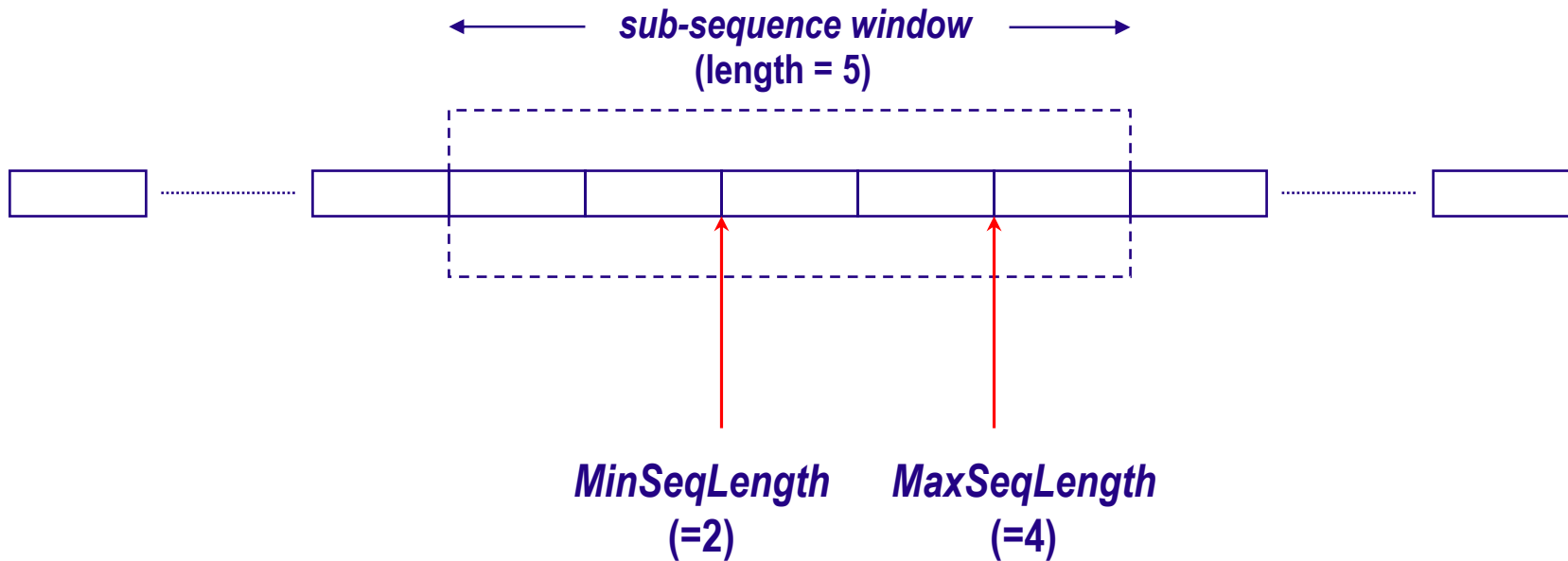
Classification performance was evaluated using the following parameters:

- Minimum and maximum sub-sequence window offset lengths,

Experiments: Performance Results

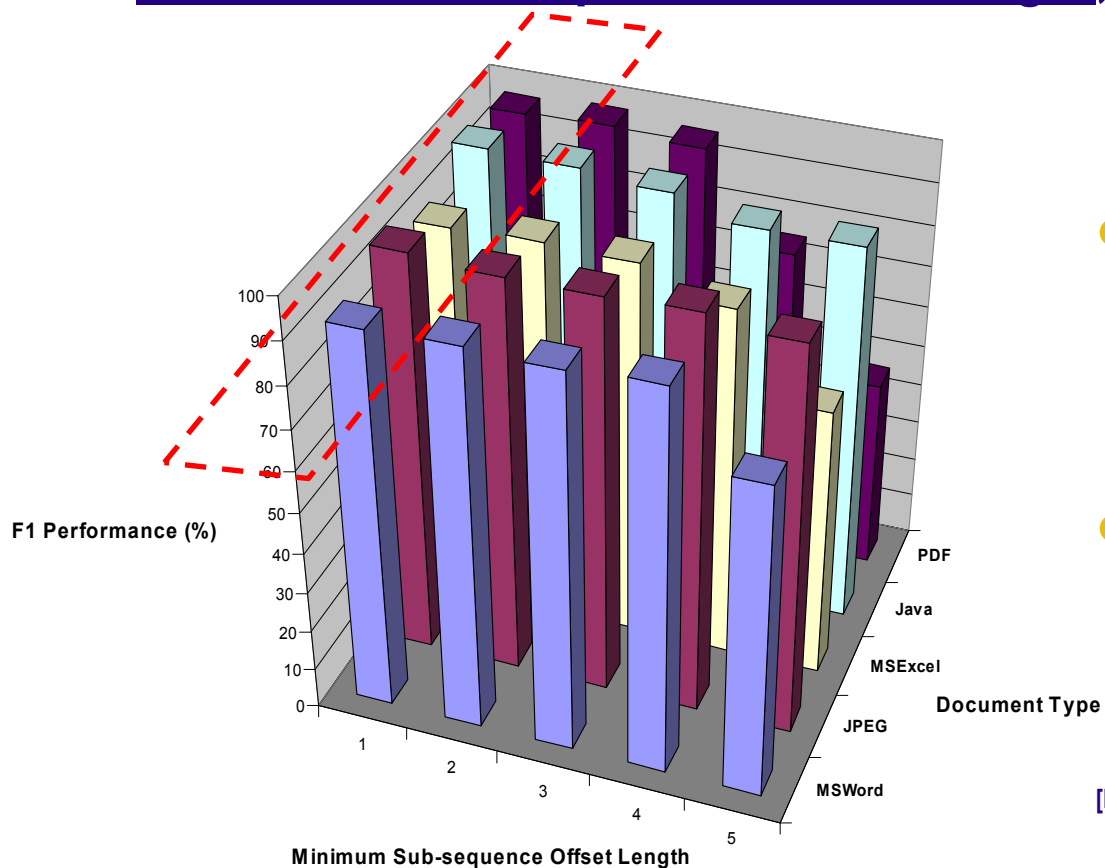
Minimum and Maximum Sub-sequence Window Offset Length:

- Definition: For example;



Experiments: Performance Results

Minimum Sub-sequence Offset Length, *MinSeqLength*:

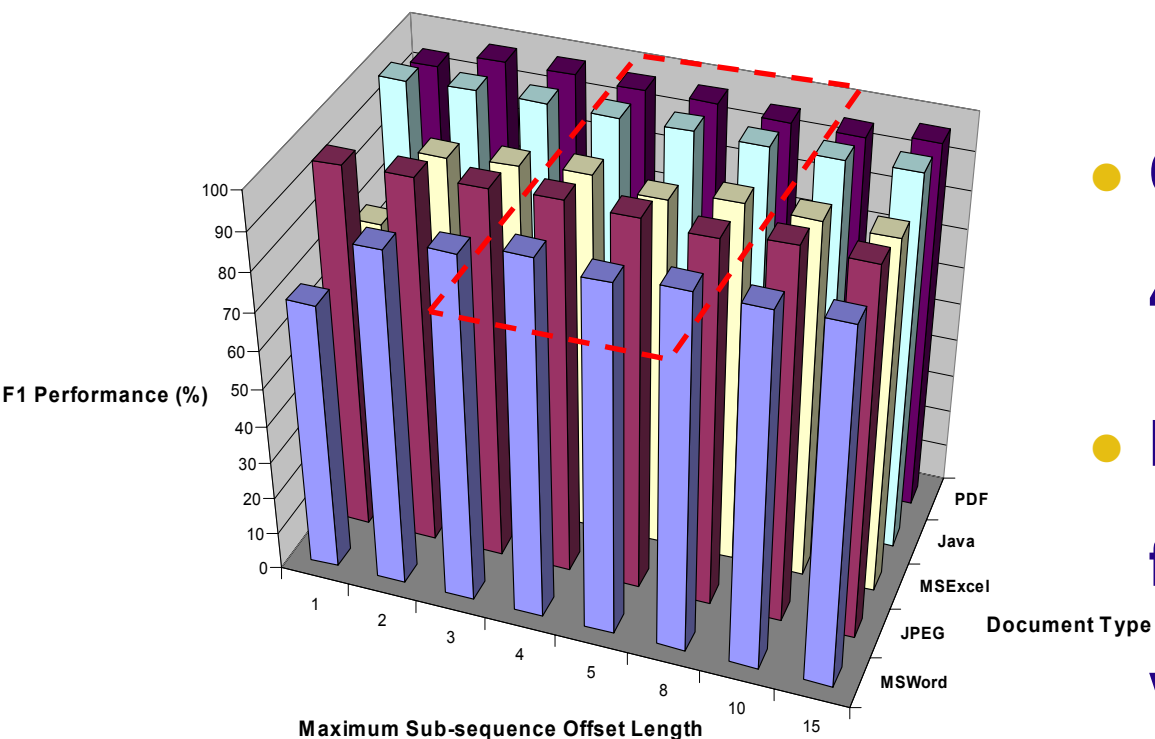


- **Optimal *MinSeqLength*=1**
 - **Use broadband spectrum**
- **Faster drop-off for non-text documents (except JPEG)**

[Parameters: *MaxStringLength*=256 *MaxSeqLength*=5 *MinSeqCount*=10]

Experiments: Performance Results

Maximum Sub-sequence Offset Length, *MaxSeqLength*:



- Optimal value is $4 < MaxSeqLength < 7$
- No significant improvements for large sub-sequence offset values

Conclusions:

- Promising semi-structured document categorization results using the broadband k -spectrum kernel.
- Experiments suggest:
 - ▶ use broadband kernels rather than narrow band kernels (ie fixed k -length sequence),
 - ▶ document length can be small,
 - ▶ use a minimum suffix count to achieve improved classification performance and tree compression.
- Extensions:
 - Increase the number of document types
 - Investigate techniques for capturing longer-range data structures in documents

Questions ?

Advertisement:

“Computer and Intrusion Forensics”

by G. Mohay, A. Anderson, B. Collie, O. de Vel and R. McKemmish
Artech House ISBN 1-58053-369-8 (2003)