DALHOUSIE UNIVERSITY
Inspiring Minds

Proximity based one-class classification with Common N-Gram dissimilarity for authorship verification task

PAN 2013 Author Identification

Magdalena Jankowska, Vlado Kešelj and Evangelos Milios

Faculty of Computer Science, Dalhousie University, Halifax, Canada

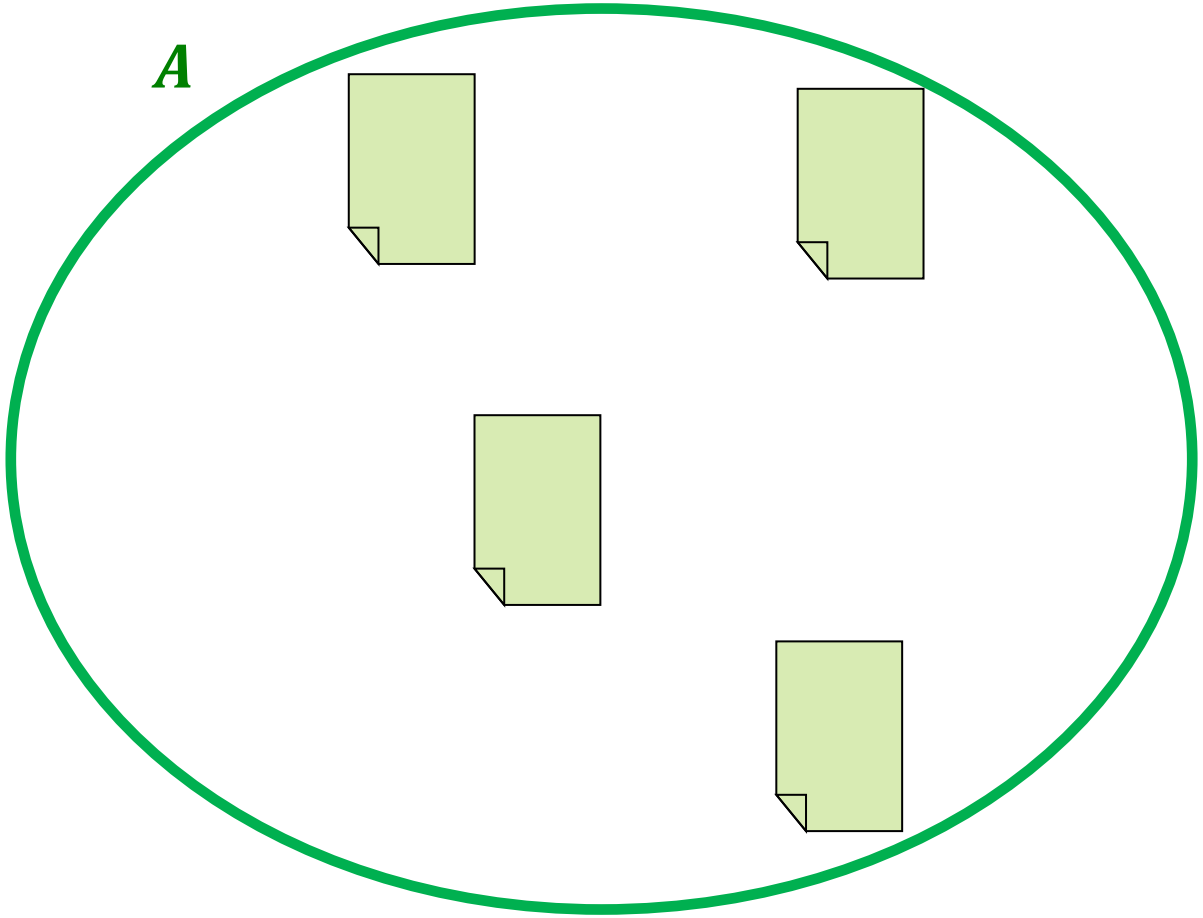
PAN Workshop, CLEF 2013, Valencia, September 25, 2013

Authorship verification problem

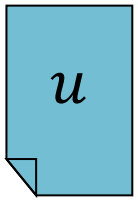
Input:

Set of "known" documents by a given author

A



document of a questioned authorship



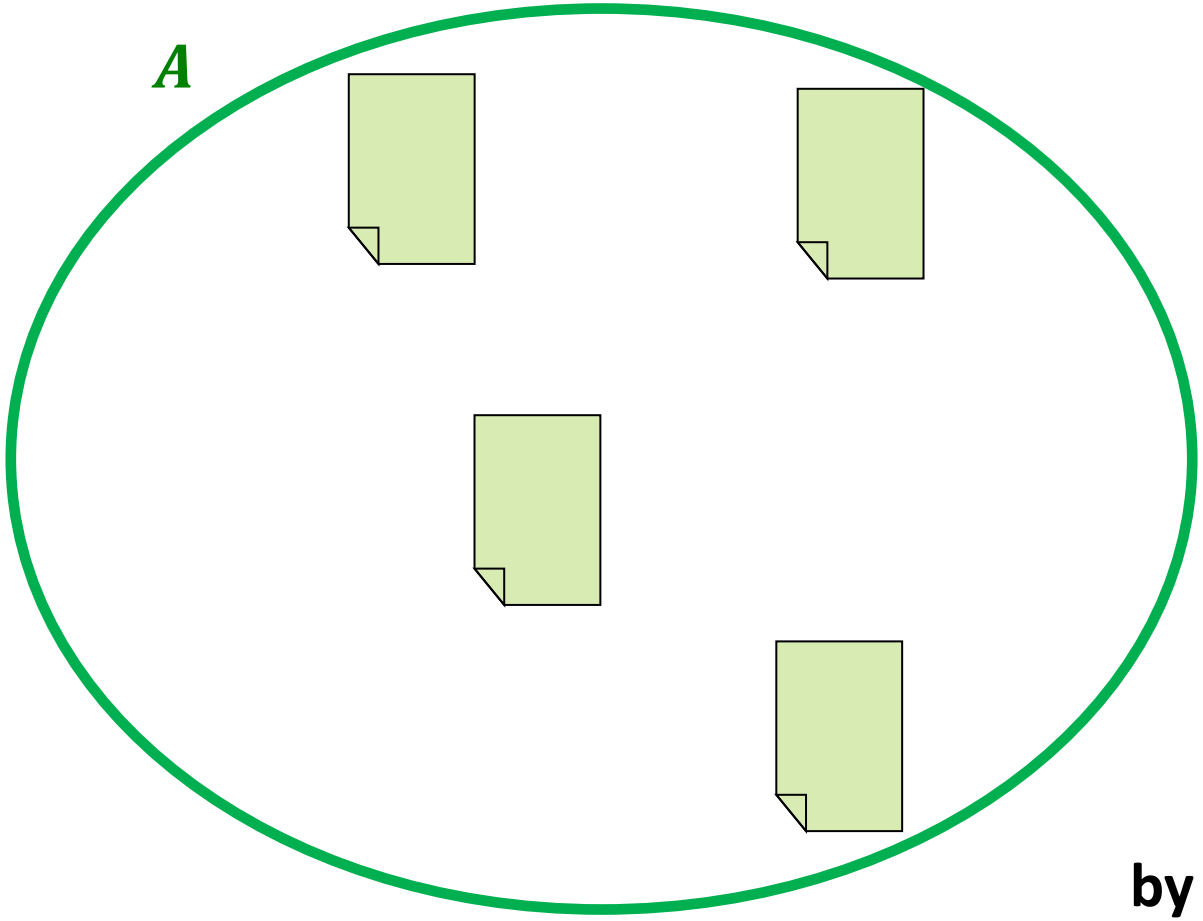
"unknown" document

Authorship verification problem

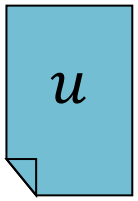
Input:

Set of "known" documents by a given author

A



document of a questioned authorship



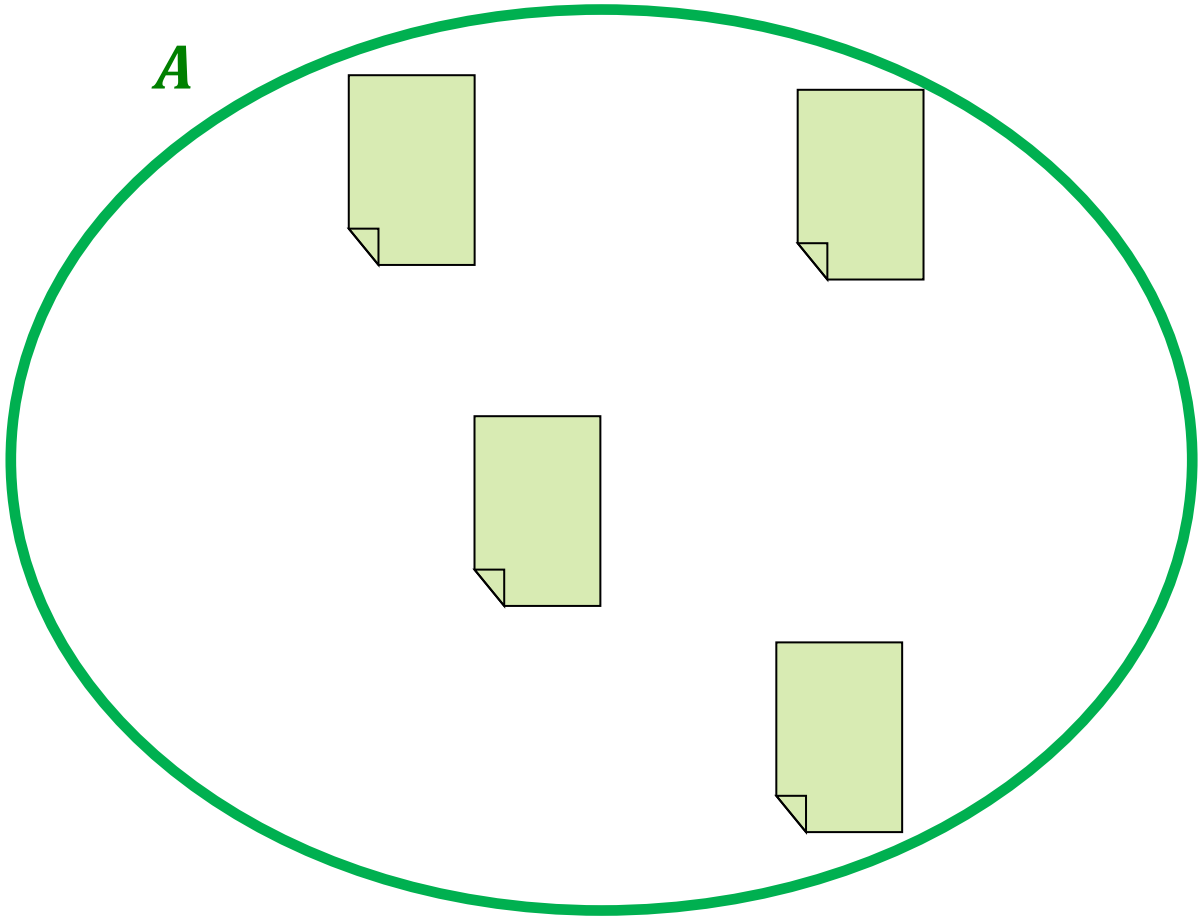
"unknown" document

Question:

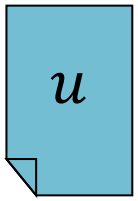
Was *u* written by the same author?

Our approach to the authorship verification problem

- **Proximity-based one-class classification**. Is ***u*** "similar enough" to ***A***?
- Idea similar to the k-centres method for one-class classification
- Applying **CNG dissimilarity** between documents



document of
a questioned authorship



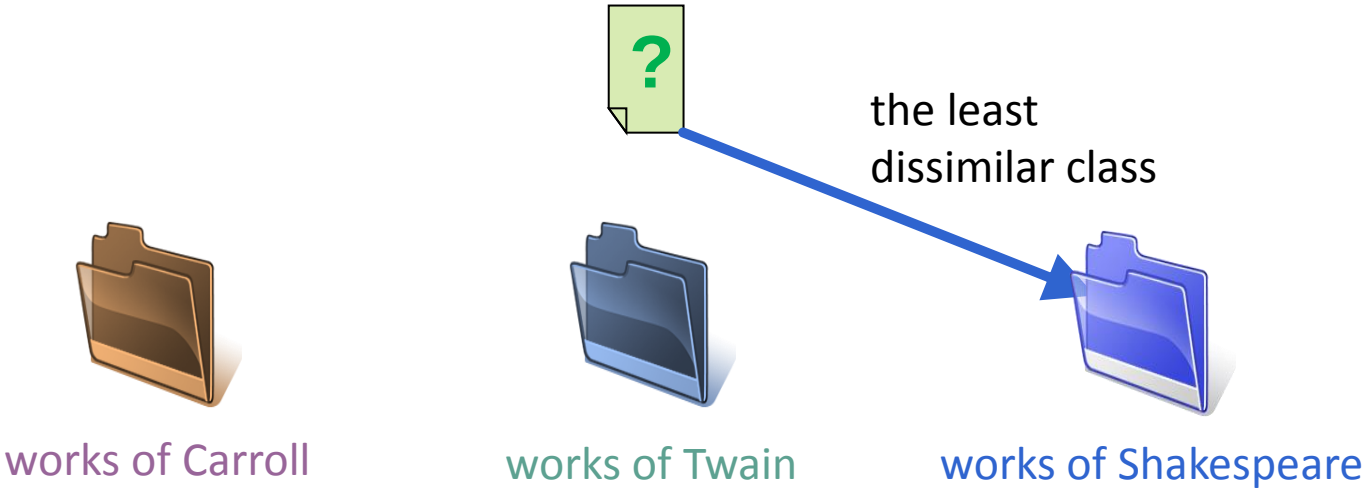
"unknown" document

Common N-Gram (CNG) dissimilarity

Proposed by

Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas.
N-gram-based author profiles for authorship attribution.
In Proc. of the Conference Pacific Association for Computational Linguistics, 2003.

Proposed as a dissimilarity measure
of the **Common N-Gram (CNG) classifier** for multi-class classification



Successfully applied to the authorship attribution problem

CNG dissimilarity - formula

Profile

a sequence of L most common n-grams of a given length n

CNG dissimilarity - formula

Profile

a sequence of L most common n-grams of a given length n

Example for **n=4, L=6**

document 1:
Alice's Adventures in the Wonderland
by Lewis Carroll

profile P_1	
n-gram	normalized frequency f_1
_ t h e	0.0127
t h e _	0.0098
a n d _	0.0052
_ a n d	0.0049
i n g _	0.0047
_ t o _	0.0044

CNG dissimilarity - formula

Profile

a sequence of L most common n-grams of a given length n

Example for **n=4, L=6**

document 1:
Alice's Adventures in the Wonderland
by Lewis Carroll

profile P_1	
n-gram	normalized frequency f_1
_ t h e	0.0127
t h e _	0.0098
a n d _	0.0052
_ a n d	0.0049
i n g _	0.0047
_ t o _	0.0044

document 2:
Tarzan of the Apes
by Edgar Rice Burroughs

profile P_2	
n-gram	normalized frequency f_2
_ t h e	0.0148
t h e _	0.0115
a n d _	0.0053
_ o f _	0.0052
_ a n d	0.0052
i n g _	0.0040

CNG dissimilarity - formula

Profile

a sequence of L most common n-grams of a given length n

Example for **n=4, L=6**

document 1:
Alice's Adventures in the Wonderland
by Lewis Carroll

document 2:
Tarzan of the Apes
by Edgar Rice Burroughs

profile P_1	
n-gram	normalized frequency $\mathbf{f_1}$
_ t h e	0.0127
t h e _	0.0098
a n d _	0.0052
_ a n d	0.0049
i n g _	0.0047
_ t o _	0.0044

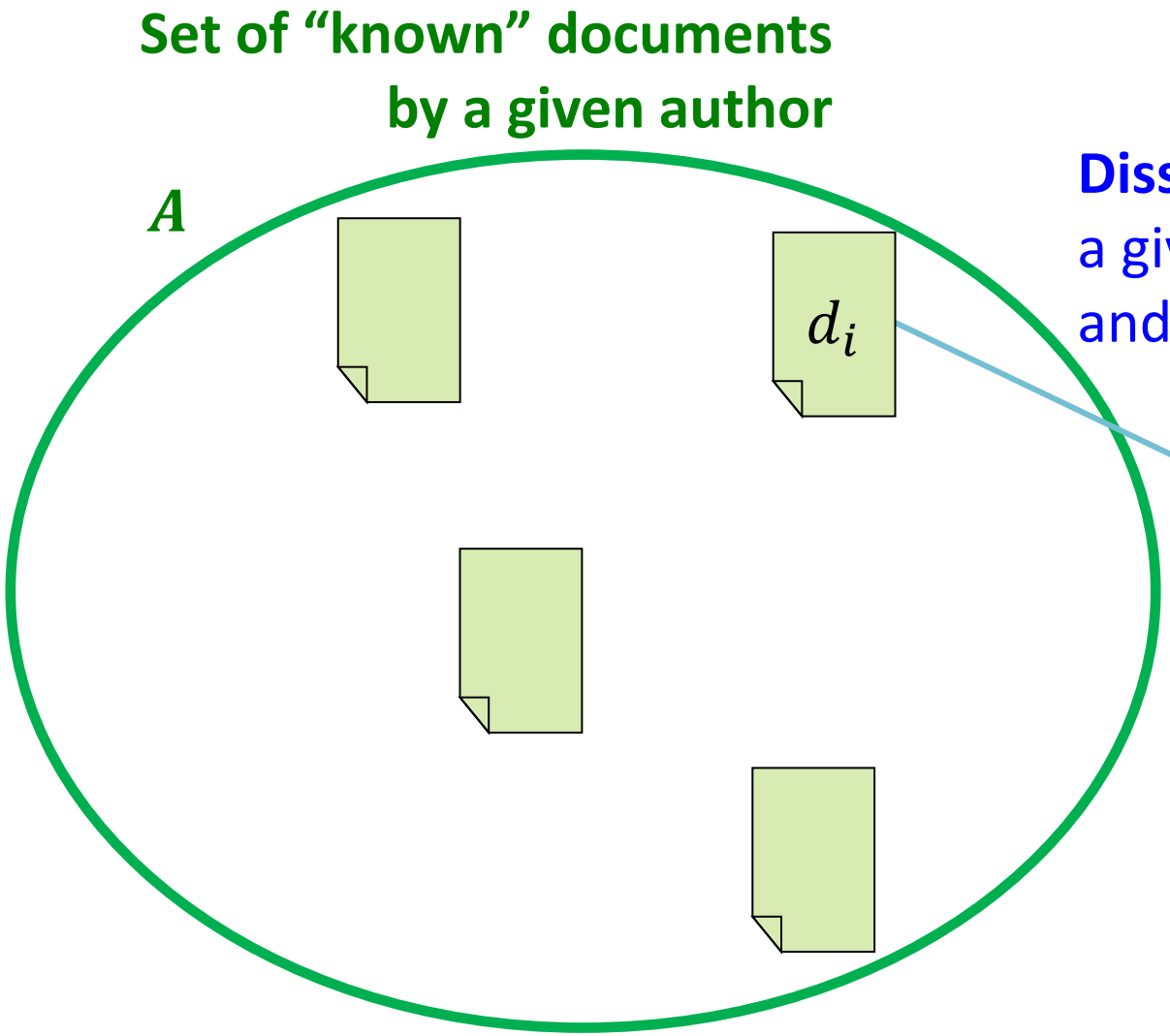
CNG dissimilarity between these documents

$$D = \sum_{x \in P_1 \cup P_2} \left(\frac{f_1(x) - f_2(x)}{\left(\frac{f_1(x) + f_2(x)}{2} \right)} \right)^2$$

where
$f_i(x) = 0$
if x does not appear in P_i

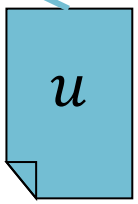
profile P_2	
n-gram	normalized frequency $\mathbf{f_2}$
_ t h e	0.0148
t h e _	0.0115
a n d _	0.0053
_ o f _	0.0052
_ a n d	0.0052
i n g _	0.0040

Proximity-based one-class classification: dissimilarity between instances



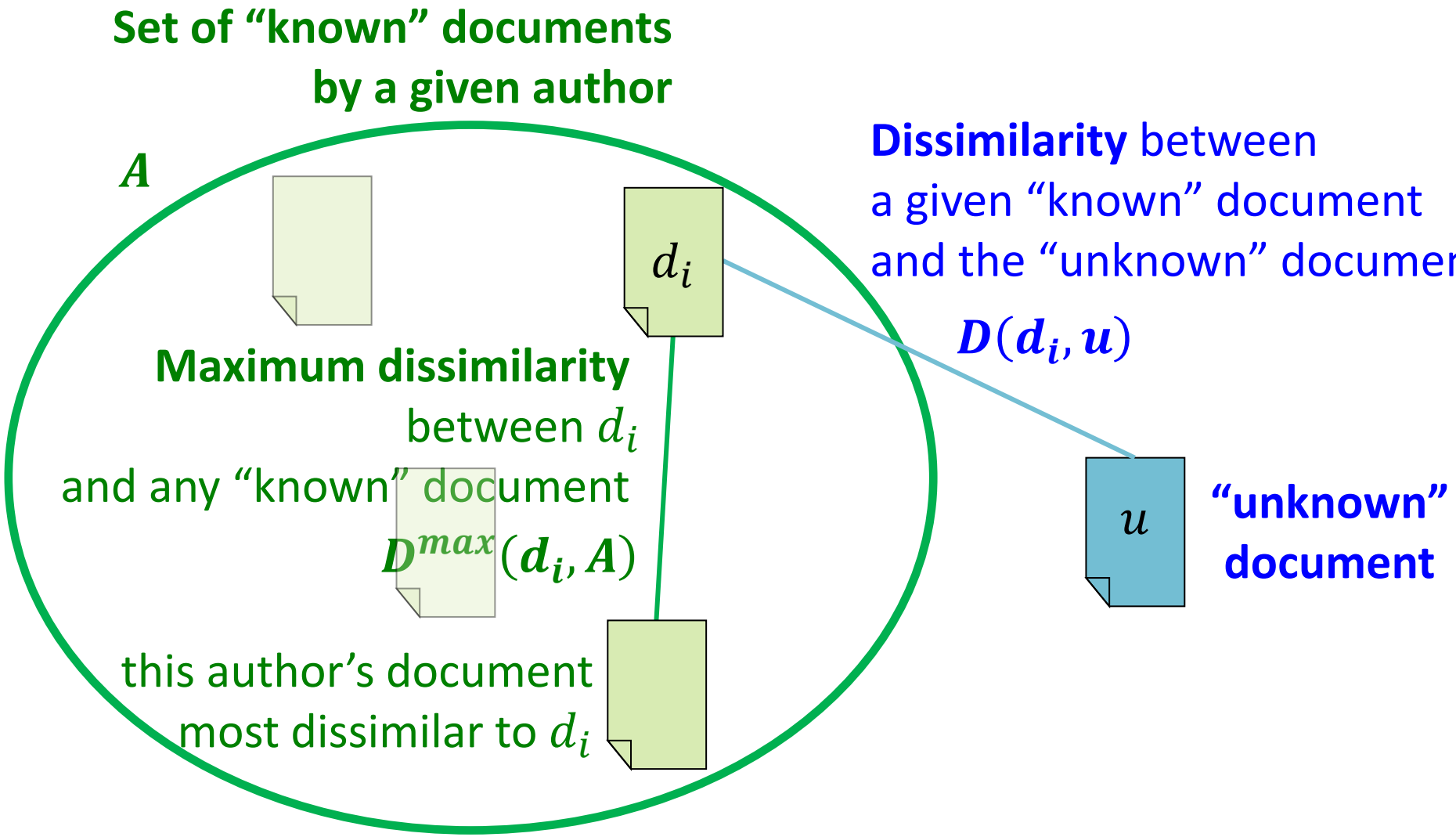
Dissimilarity between
a given “known” document
and the “unknown” document

$\boldsymbol{D(d_i, u)}$



“unknown” document

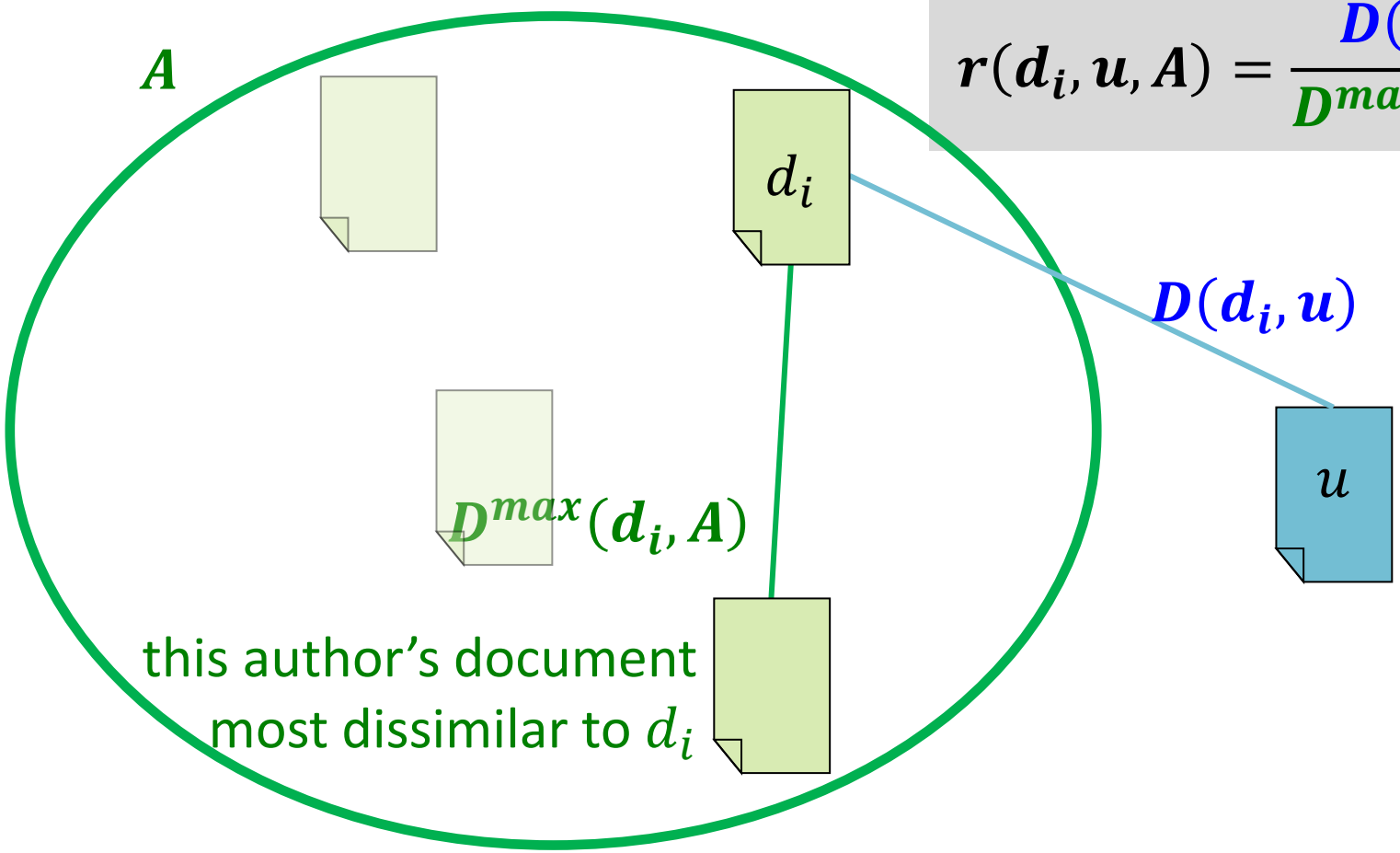
Proximity-based one-class classification: dissimilarity between instances



Proximity-based one-class classification: dissimilarity between instances

Dissimilarity ratio of d_i :
How much more/less dissimilar is the "unknown" document than the most dissimilar document by the same author.

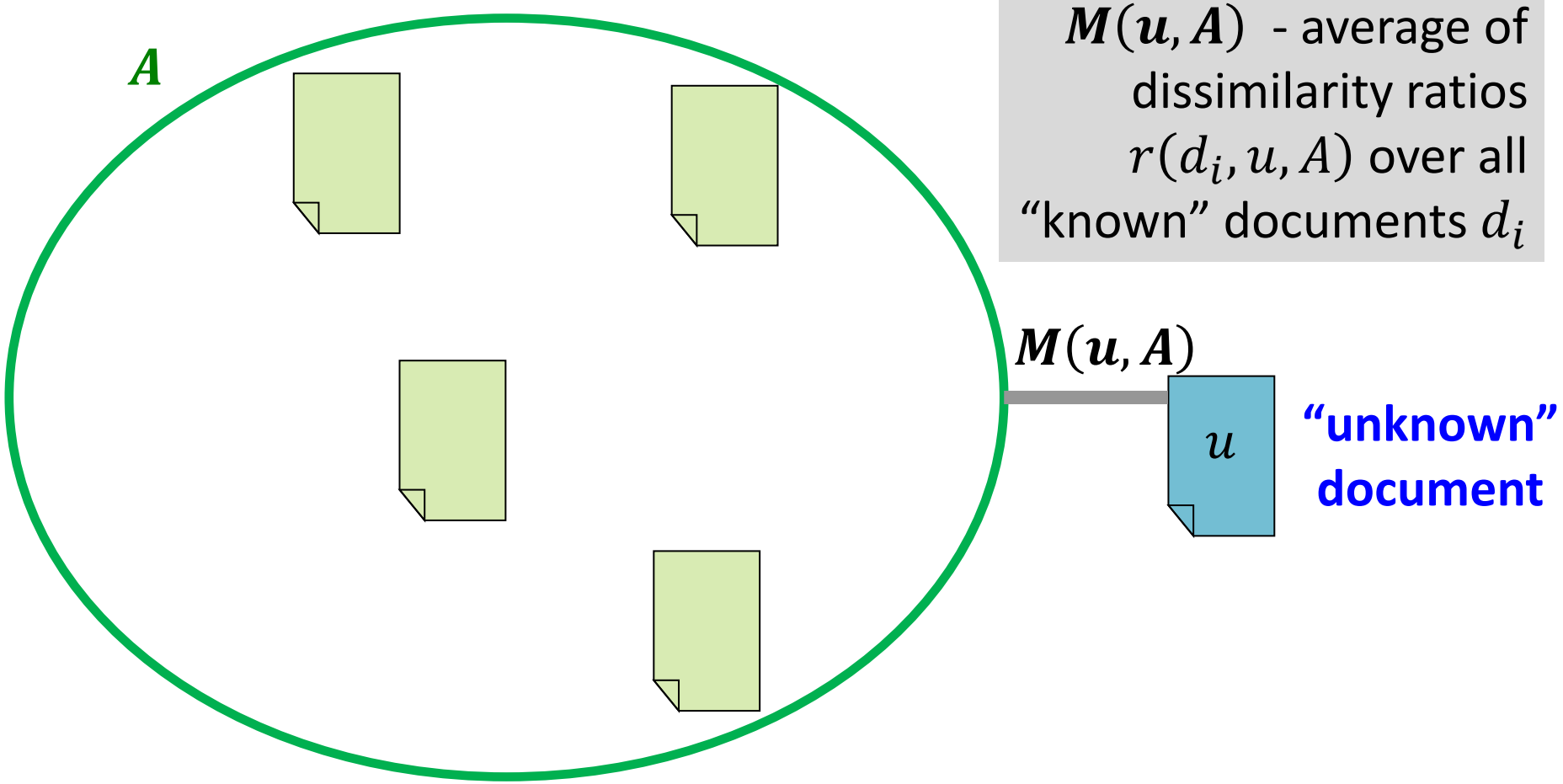
$$r(d_i, u, A) = \frac{D(d_i, u)}{D^{max}(d_i, A)}$$



Proximity-based one-class classification: proximity between a sample and the positive class instances

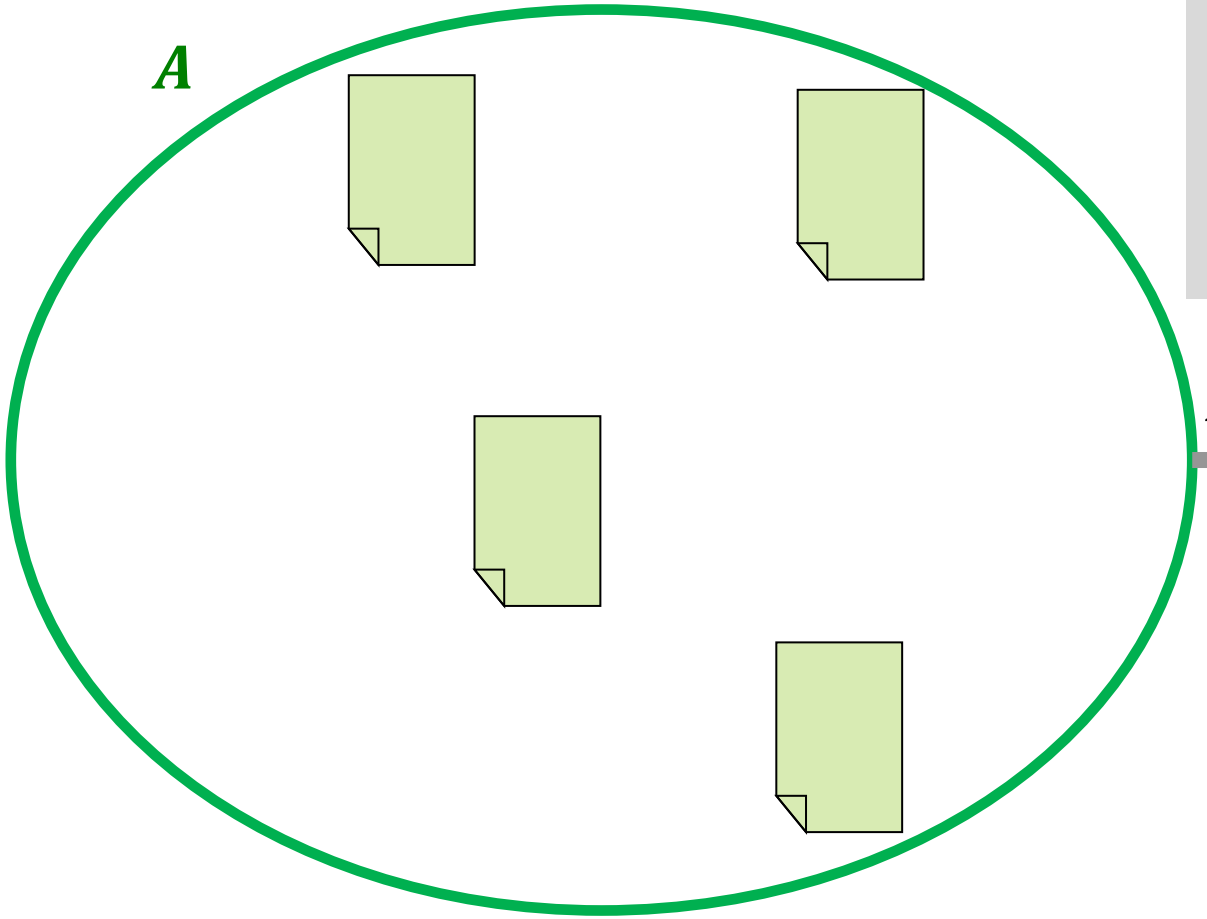
Measure of proximity between the "unknown" document and the set *A* of documents by a given author:

$\boldsymbol{M(u, A)}$ - average of dissimilarity ratios $r(d_i, u, A)$ over all "known" documents d_i



Proximity-based one-class classification: thresholding on the proximity

Iff $\boldsymbol{M(u, A)}$ less than or equal to a threshold θ :
classify *u* as belonging to *A*
i.e., written by the same author



$\boldsymbol{M(u, A)}$ - average of dissimilarity ratios $r(d_i, u, A)$ over all "known" documents d_i

$\boldsymbol{M(u, A)}$

u

"unknown" document

Real scores

Obtained by linear scaling the $M(u, A)$ measure:

the threshold θ → 0.5

with **cut-off** at $\theta \pm 0.1$:

$M(u, A) < \theta - 0.1$ → 1

$M(u, A) > \theta + 0.1$ → 0

Special conditions used

- Dealing with instances **when only 1 "known" document** by a given author is provided:
  dividing the single "known" document into two halves and treating them as two "known" documents

- Dealing with instances when some documents **do not have enough character n-grams to create a profile of a chosen length**:
  representing all documents in the instance by equal profiles of the maximum length for which it is possible

- **Additional preprocessing** (tends to increase accuracy on training data):
  cutting all documents in a given instance to an equal length in words

Parameters

Parameters of our method:

Type of tokens: we used characters

n – n-gram length
L – profile length

θ – **threshold for the proximity measure M for classification**
(biggest problem)

Parameter selection

Parameters for the final competition run selected using experiments on training data in Greek and English:

- provided by the competition organizers
- compiled by ourselves from existing datasets for other authorship attribution problems

For Spanish: the same parameters as for English

	English Spanish	Greek
n (length of character n-grams)	6	7
L (profile length)	2000	2000
θ (threshold) if at least two “known” documents given	1.02	1.008
θ (threshold) if only one “known” document given	1.06	1.04

Results on PAN 2013 competition test dataset

	Entire set	English subset	Greek subset	Spanish subset
F_1 of our method	0.659	0.733	0.600	0.640
competition rank	5th (shared) of 18	5th (shared) of 18	7th (shared) of 16	9th of 16
best F_1 of other competitors	**0.753**	**0.800**	**0.833**	**0.840**
AOC	**0.777**	**0.842**	0.711	0.804

Conclusion

- Very encouraging results in terms of the power of our measure M for ordering the instances
- Difficult choice of the threshold, depending much on the corpus

Future work

- Further parameter analysis
- Exploration of involving a user interaction and insight through visualization
- Exploration of improvements of the method

Acknowledgement

- This research was funded by a contract from the Boeing Company, a Collaborative Research and Development grant from the Natural Sciences and Engineering Research Council of Canada, and Killam Predoctoral Scholarship.

Thank you!