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

Magdalena Jankowska, Vlado Kešelj and Evangelos Milios

Faculty of Computer Science, Dalhousie University

September 2014

Example

"The Cuckoo's Calling“

2013 detective novel by Robert Galbraith

Example

"The Cuckoo's Calling“

2013 detective novel by Robert Galbraith

Question by *Sundays Times*

Was “The Cuckoo’s Calling” really written by J.K. Rowling?

⬇

Peter Millican and Patrick Juola requested (independently) to answer this question through their algorithmic methods

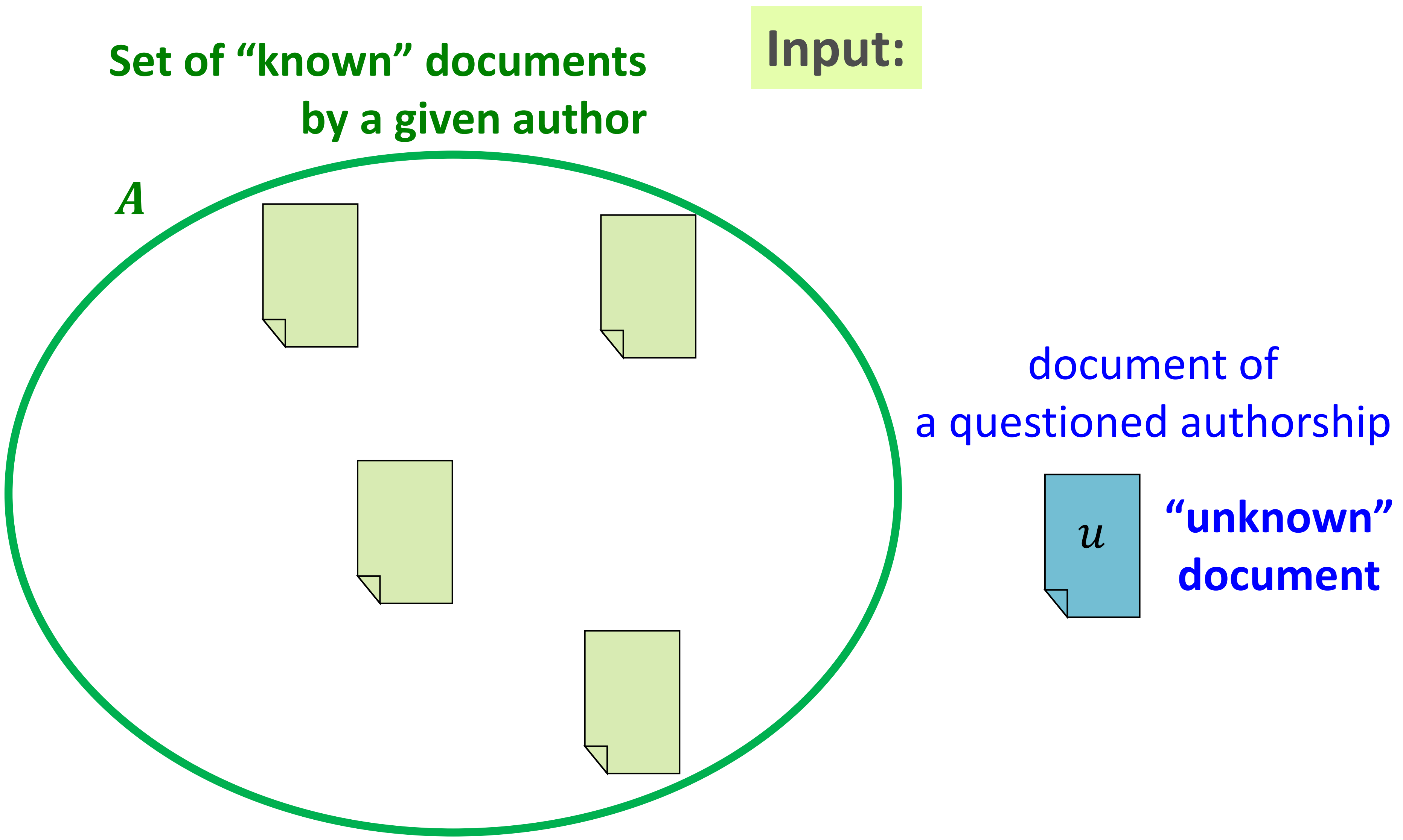
⬇

Results indicative of the positive answer

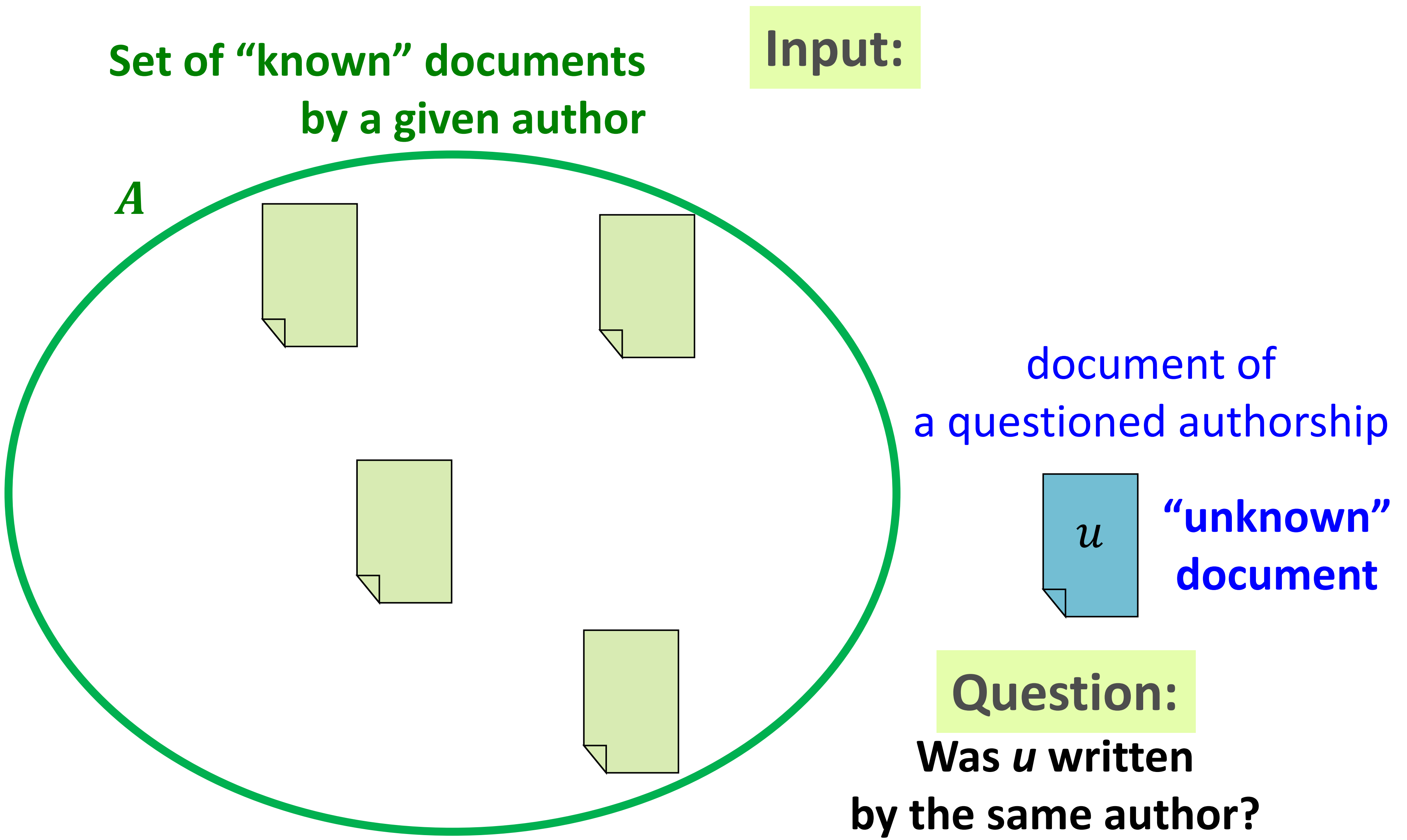
⬇

J. K. Rowling admitted that she is the author

Authorship verification problem



Authorship verification problem



Motivation

Applications:

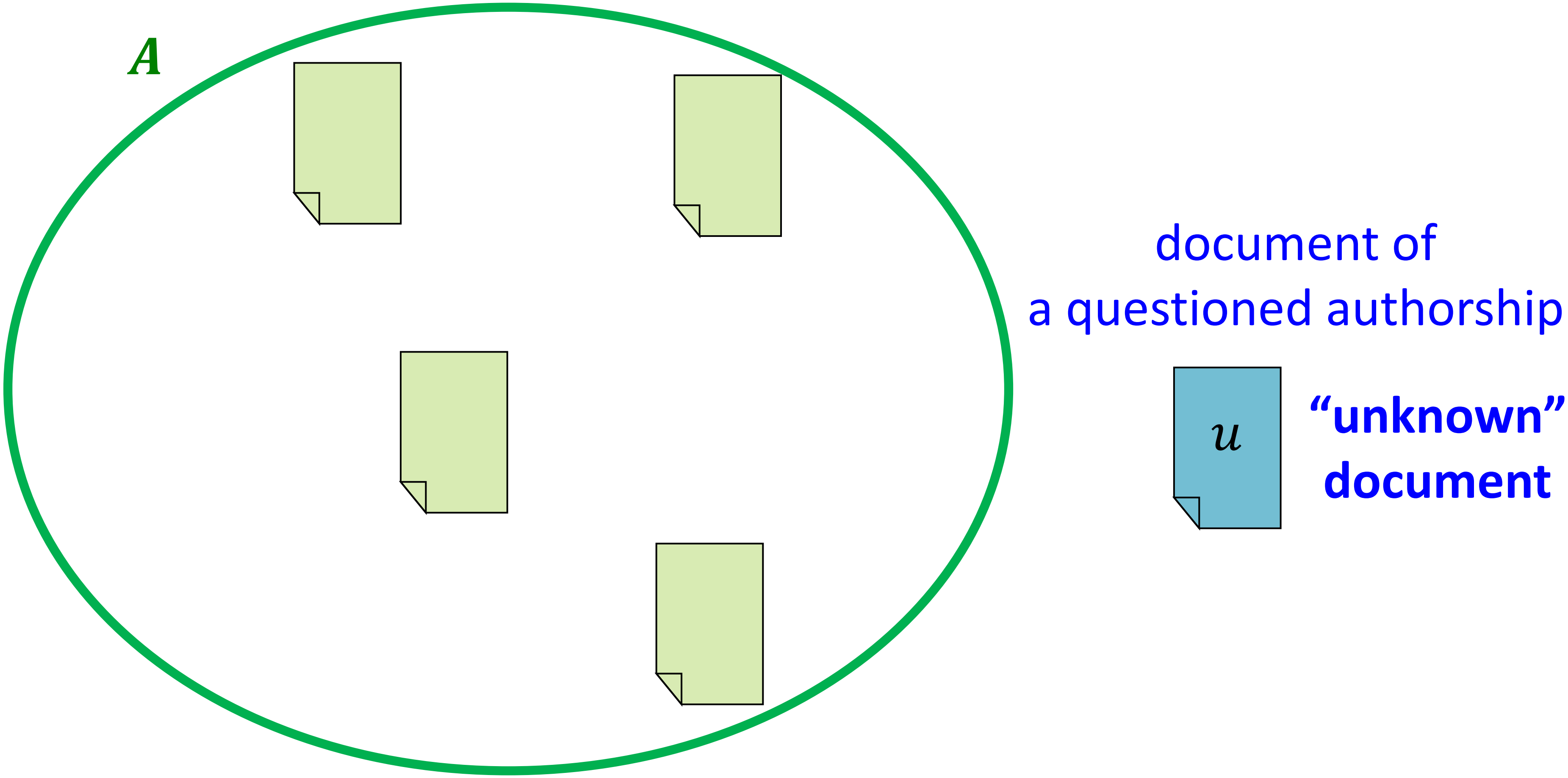
Forensics

Security

Literary research

Our approach to the authorship verification problem

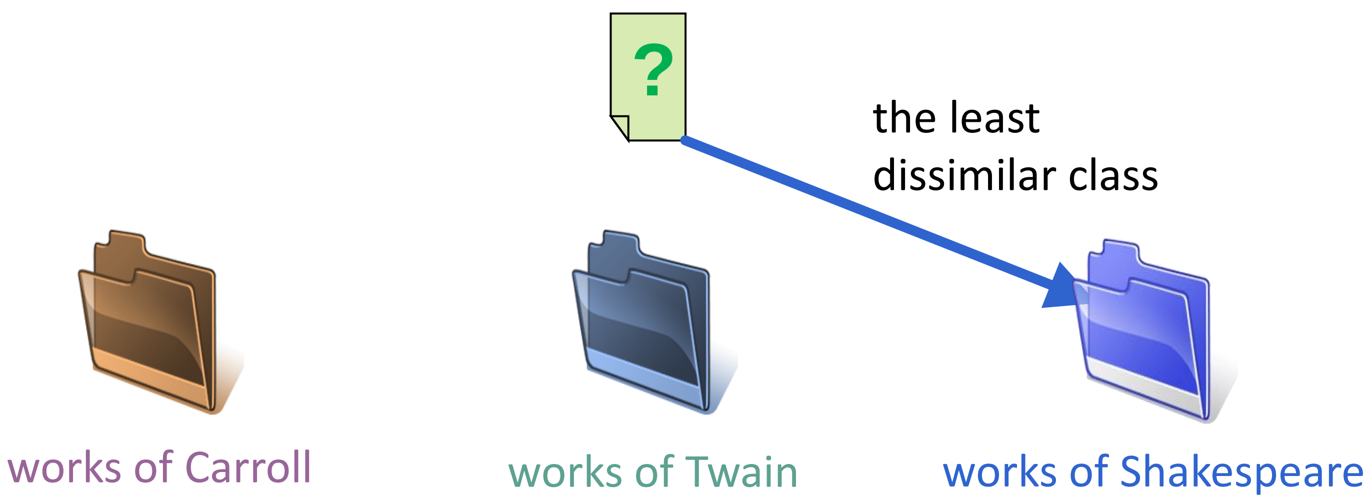
- **Proximity-based one-class classification**. Is ***u*** "similar enough" to ***A***?
- Idea similar to the k-centres method for one-class classification
- Applying **CNG dissimilarity** between documents



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

Character n-grams

Strings of n consecutive characters from a given text

$\boxed{\text{Alic}}$e was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

n=4
4-grams

Alic

Character n-grams

Strings of n consecutive characters from a given text

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

n=4
4-grams

Alic
lice

Character n-grams

Strings of n consecutive characters from a given text

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

n=4
4-grams

Alic
lice
ice_

Character n-grams

Strings of n consecutive characters from a given text

Ali$\boxed{\text{ce w}}$as beginning to get very tired of sitting by her sister on the bank, and of having nothing to do:

Alice's Adventures in the Wonderland
by Lewis Carroll

n=4
4-grams

Alic
lice
ice_
ce_w

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 $\mathbf{f_1}$
_ 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 - 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

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

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

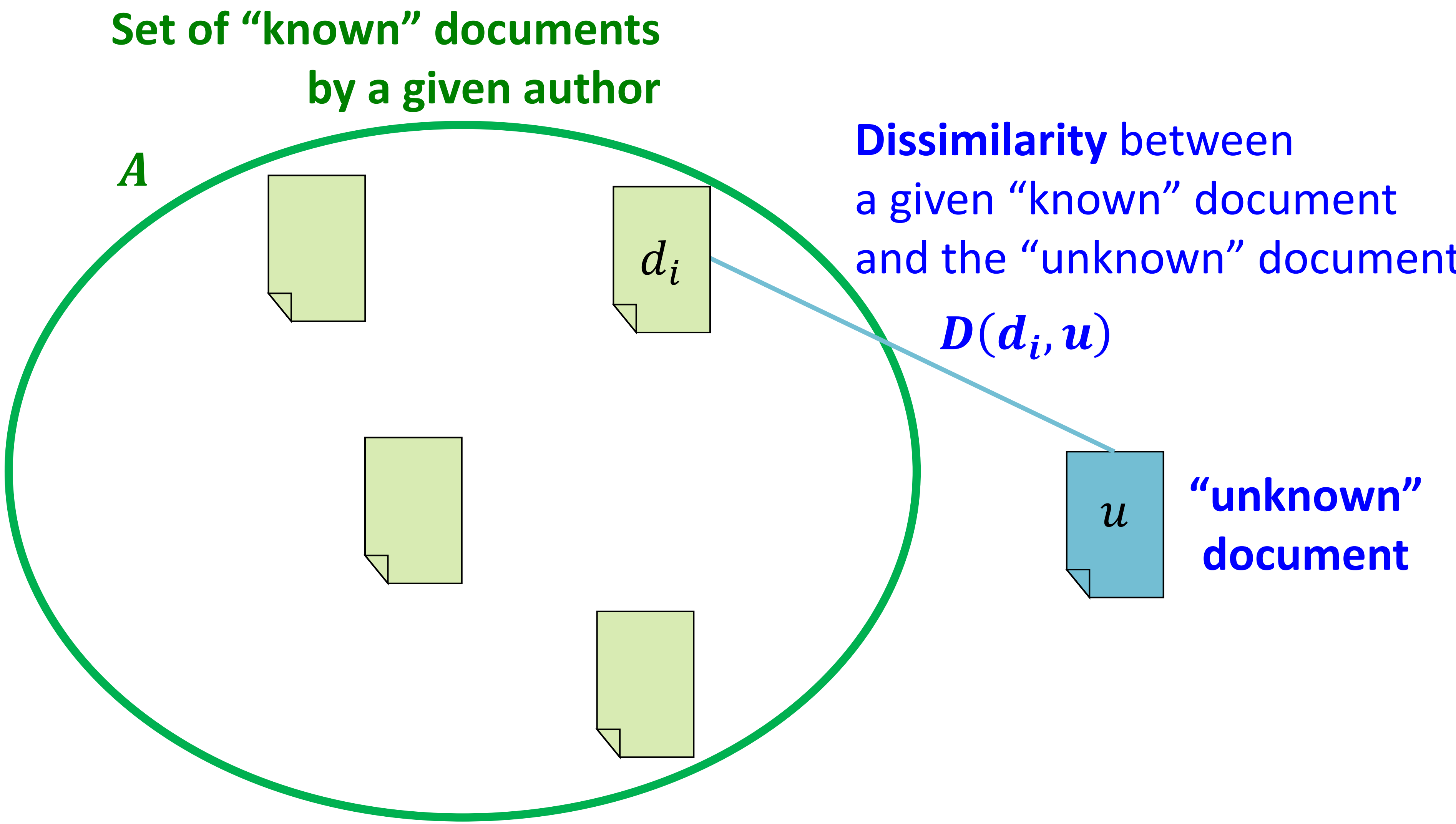
CNG dissimilarity between these documents

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

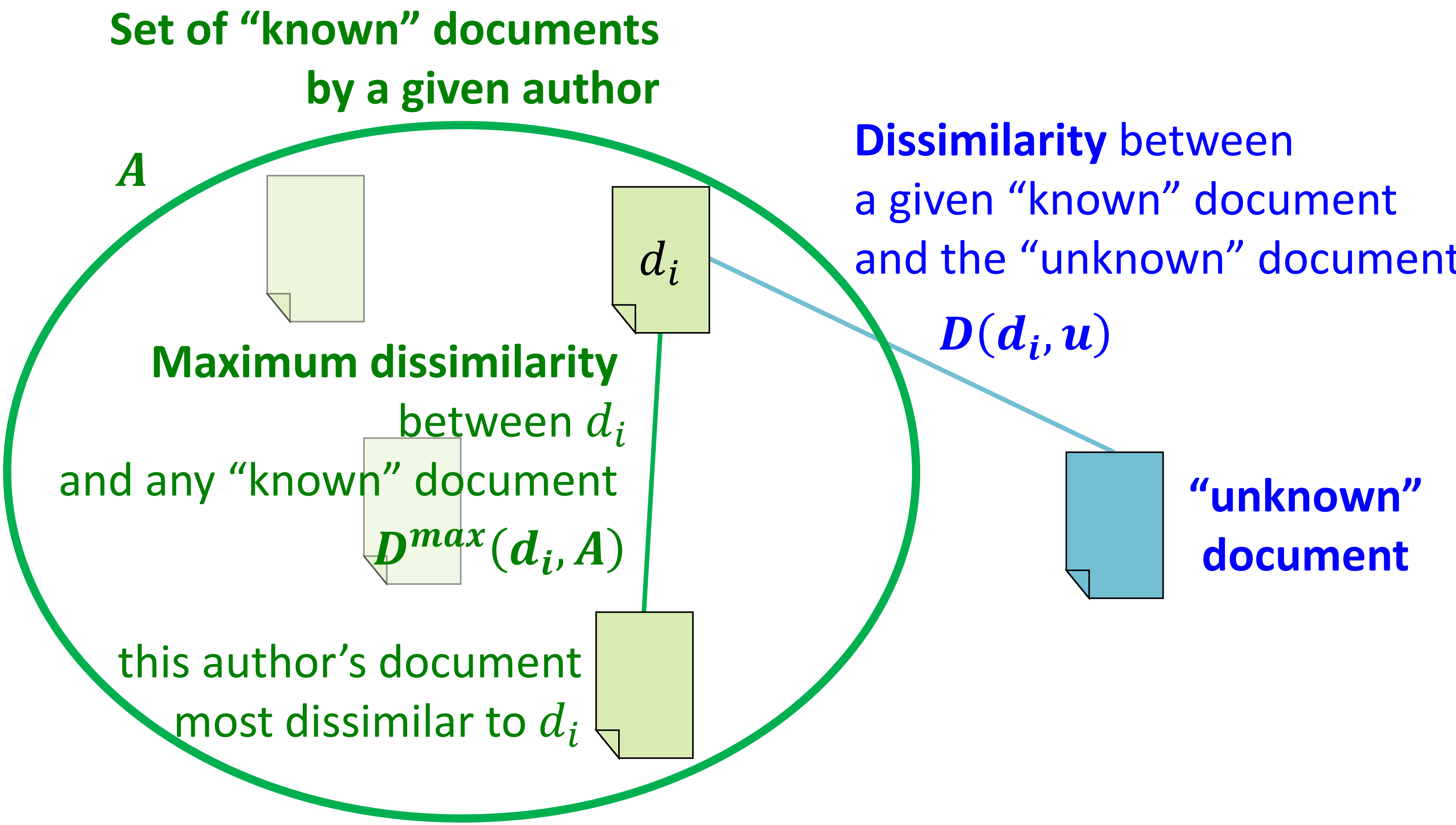
where
$f_i(x) = 0$
if x does not appear in P_i

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



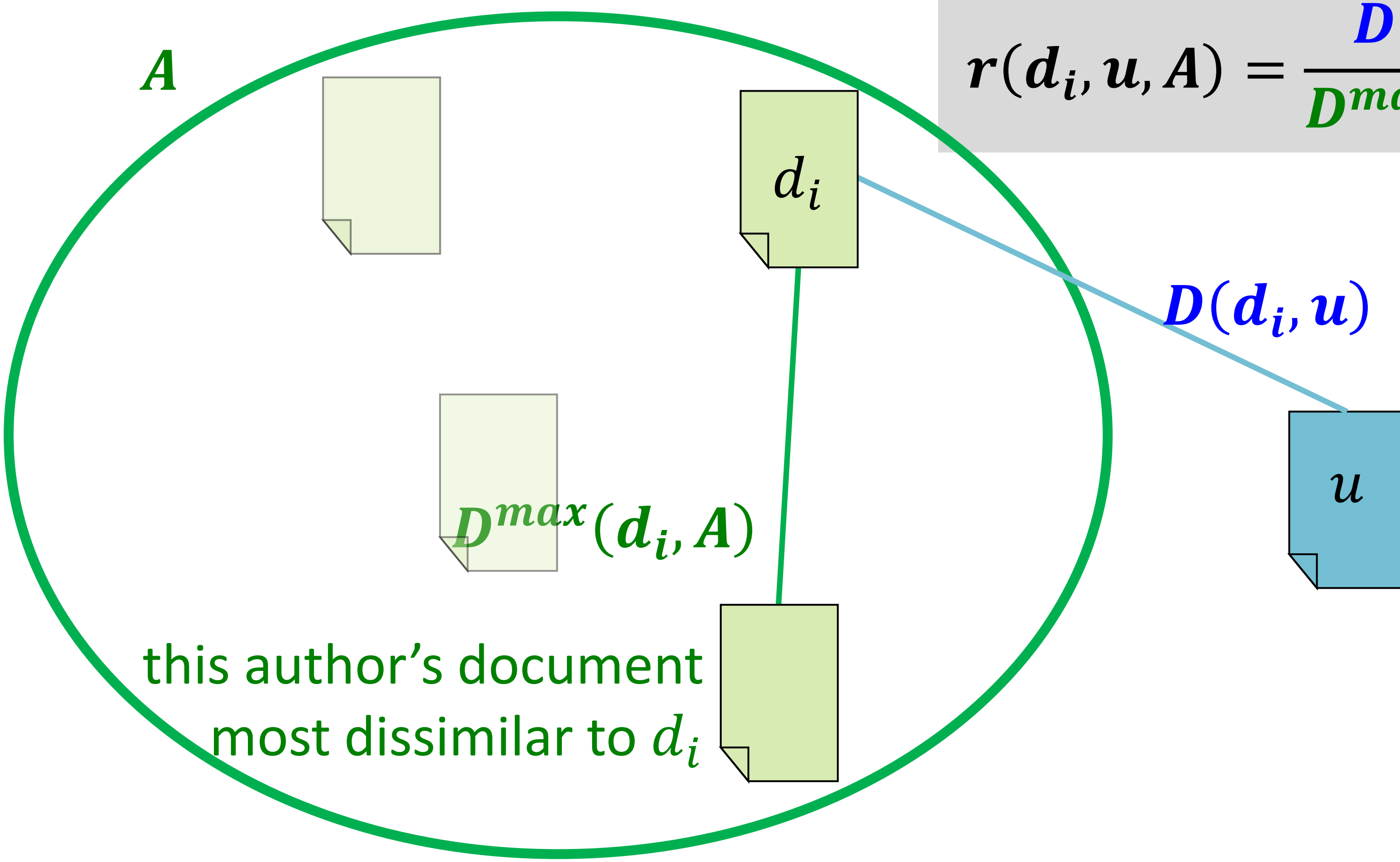
Proximity-based one-class classification: dissimilarity between instances



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.

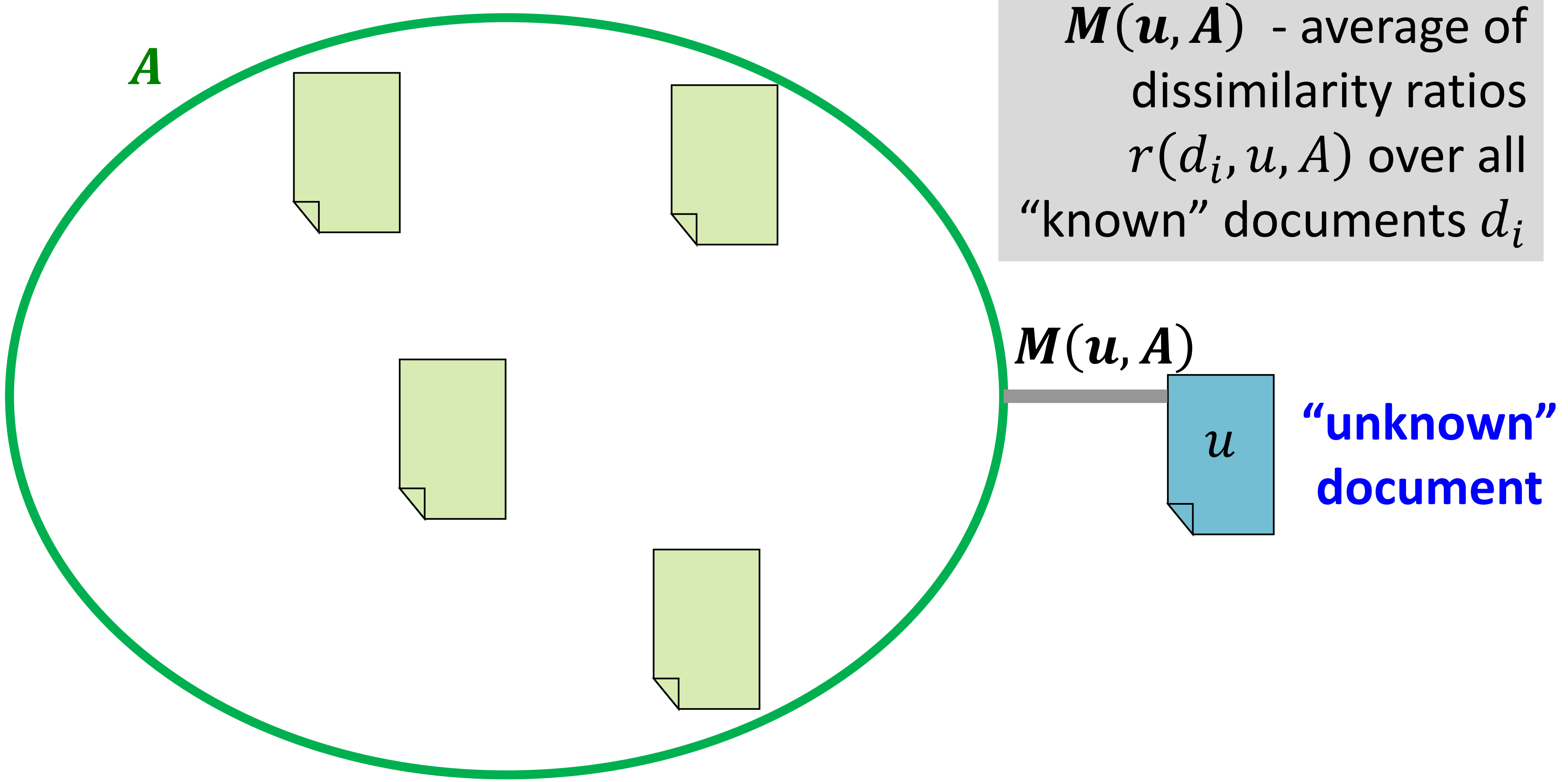
$$r(d_i, u, A) = \frac{D(d_i, u)}{D^{max}(d_i, A)}$$



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

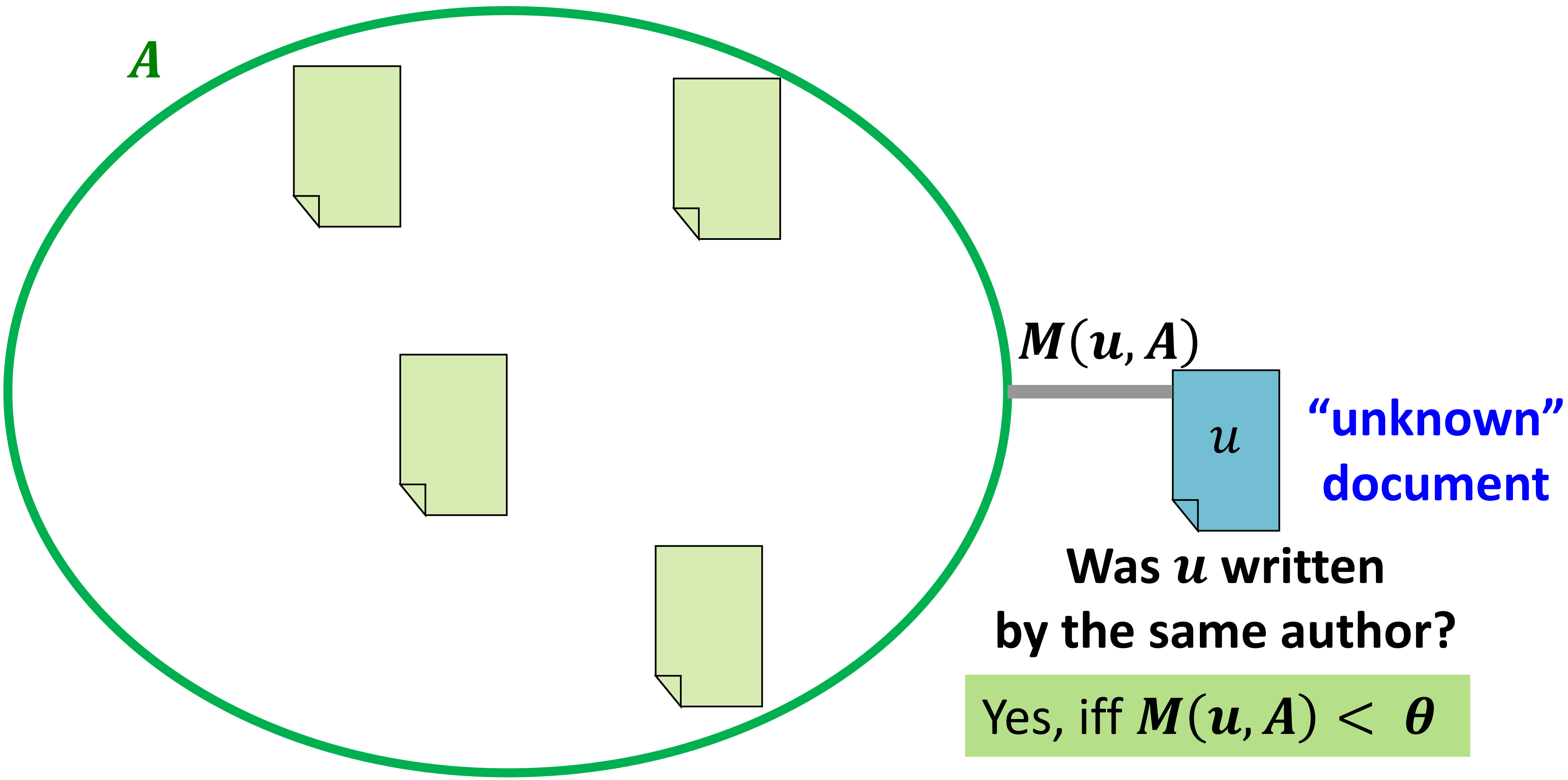
Measure of proximity between the "unknown" document and the set A of documents by a given author:

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



Was u written by the same author?

Yes, iff $\boldsymbol{M(u, A) < \theta}$

Confidence scores

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

the threshold θ → 0.5

with a **cut-off** α at $\theta \pm$ α :
$M(u, A) < \theta -$ α → 1
$M(u, A) > \theta +$ α → 0

Parameters

Parameters of our method:

n – n-gram length
L – profile length: number of the most common n-grams considered
θ – **threshold for the proximity measure M for classification**

Proximity-based one-class classification: 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

Ensembles

Ensembles combine classifiers that differ between each other with respect to at least one of the three document representation parameters:

- type of the tokens for n-grams (word or character)
- size of n-grams
- length of a profile

Aggregating results by majority voting or voting weighted by confidence score by a classifier

Testbed1: PAN 2013 competition dataset

PAN 2013 – 9th evaluation lab on
uncovering plagiarism, authorship, and social software misuse

Author Identification task:
Author Verification problem instances in English, Greek and Spanish

Competition submission: a single classifier

Parameters for the competition submission 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 (n-gram length)	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 of PAN 2013 competition submission

	Entire set	English subset	Greek subset	Spanish subset
Evaluation measure: F_1 (identical to accuracy for our method)				(18 teams)
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**
Secondary competition evaluation measure: AUC (10 teams)				
AUC	**0.777**	**0.842**	0.711	0.804
competition rank	1st of 10	1st of 10	2nd of 10	2nd of 10
Best AUC of 9 other participants	0.735	0.837	**0.824**	**0.926**

Evaluation of ensembles on PAN 2013 dataset (after contest)

Selected experimental results for ensembles	**Entire set**		**English subset**		**Spanish subset**		**Greek subset**	
	Acc.	AUC	Acc.	AUC	Acc.	AUC	Acc.	AUC
Our ensembles: weighted voting, all classifiers in the considered parameter space								
character based	0.729	0.764	**0.833**	0.830	0.800	0.859	0.567	0.582
character and word based	0.741	0.780	0.800	0.842	**0.840**	0.853	0.600	0.622
Our ensemble: weighted voting, classifiers selected based on performance on train data								
character and word based	**0.788**	**0.805**	0.800	**0.857**	**0.840**	0.853	0.733	0.687
Methods by other PAN'13 participants (different methods in different columns)								
best results over other participants	0.753	0.735	0.800	0.837	**0.840**	**0.926**	**0.833**	**0.824**

Evaluation on PAN 2014 author verification competition

Difference in dataset as compared to PAN 2013:

- fewer known documents per problem (max 5), in particular two datasets where only one known document is given per problem
- more problems in testing and training set
- more data categories:
  languages: English, Dutch, Spanish, Greek
  different genre categories

Our submission to PAN 2014 competition

- Separate ensemble for each category (language&genre combination)
- Ensembles selected based on performance on training data: fixed odd number of 31 classifiers with the best AUC
- Threshold set to the average of optimal thresholds of the selected classifiers on the train data (thresholds on which maximum accuracy achieved)

Results on PAN 2014 dataset: articles in Greek and Spanish

Greek articles

our competition rank: 5th of 13

	Product of AUC and c@1	AUC	c@1
our submission	0.497	0.731	0.680
result of the top participant	0.720	0.889	0.810

Spanish articles

our competition rank: 3rd of 13

	Product of AUC and c@1	AUC	c@1
our submission	0.586	0.803	0.730
result of the top participant	0.698	0.898	0.778

Results on PAN 2014 dataset: Dutch essays and reviews

Dutch essays			
our competition rank: 6th of 13			
	Product of AUC and c@1	AUC	c@1
our submission	0.732	0.869	0.842
result of the top participant	0.823	0.932	0.883

Dutch reviews			
our competition rank: 5th of 13			
	Product of AUC and c@1	AUC	c@1
our submission	0.357	0.638	0.560
result of the top participant	0.525	0.757	0.694

Results on PAN 2014 dataset: English essays and novels

English essays

our competition rank: 12th of 13

	Product of AUC and c@1	AUC	c@1
our submission	0.284	0.518	0.548
result of the top participant	0.513	0.723	0.710

English novels

our competition rank: 13th of 13

	Product of AUC and c@1	AUC	c@1
our submission	0.225	0.491	0.457
result of the top participant	0.508	0.711	0.715

Results on PAN 2014 dataset: entire data set

PAN 2014 entire data set			
our competition rank: 9th of 13			
	Product of AUC and c@1	AUC	c@1
our submission	0.367	0.609	0.602
result of the top participant	0.490	0.718	0.683

Discussion of results on PAN 2013 and PAN 2014 datasets

The ensembles of word-based and character based classifiers with weighted voting and that used the training data were tested on both PAN 2013 and PAN 2014 sets

- Our method is best suited for problems with at least 3 “known” documents (as it takes advantage of the pair of the most dissimilar known documents). On all evaluation sets in which the average number of known documents is at least 3 per problem, the results were satisfactory (corresponding to the 3^{rd} or higher competition rank):
  - PAN 2013 entire set
  - PAN 2013 English set
  - PAN 2013 Spanish set
  - PAN 2013 Greek set
  - PAN 2014 Spanish articles set
- Problems with only one known documents are very challenging for our method. On the two datasets for which the number of known documents was 1 per problem, the results were very poor:
  - PAN 2014 English novels
  - PAN 2014 Dutch reviews
- More investigation is needed for explaining the extremely poor performance on PAN 2014 English essays. One special feature of this set is that is the only one where the authors are not native speakers

Conclusion

An intrinsic one-class proximity based classification for authorship verification

Evaluated on datasets of PAN 2013 and PAN 2014 author verification competition: competitive results for sets with the average number of documents of known authorship is at least 3

Poor results on problems with only 1 document of known authorship

Ensembles of character based and word based classifiers seems to work best

Future work

- Better adaptation of the method for the problems where only one known document is present
  - Investigating dividing known documents into more chunks instead of just two. This may also be applied and possibly improve the performance for cases when 2 known documents are present

- analysis of the role of word n-grams and character n-grams depending on the genre of the texts, and on the topical similarity between the documents

Thank you!