Authorship Verification, combining linguistic features and different similarity functions

Daniel Castro Castro[1], Yaritza Adame Arcia[1], María Pelaez Brioso[1], Rafael Muñoz Guillena[2]
[1]Desarrollo de Aplicaciones, Tecnología y Sistemas
DATYS, Cuba
{daniel.castro, yaritza.adame, maria.pelaez}@datys.cu
[2]Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante, España
rafael@dlsi.ua.es

Abstract

We propose an authorship analysis method that compares the **average similarity** of a text of **unknown** authorship with **all the texts of an author**. Using this idea, a text that was not written by an author, would not exceed the average of similarity with known texts and a text of unknown authorship would be considered as written by the author, only if it exceeds the average of similarity obtained between texts written by him and if it got the major value comparing the average similarity with the rest of the authors.

Contribution

- The idea of the **AGS** measure as a **limit** to determine when an unknown document was written by an author. This could be a strict limit to determine when a text was written by an author.
- Take a **final-decision** based on the combination of the results of pair function-feature for each linguistic feature, and all the decisions using the total number of features.

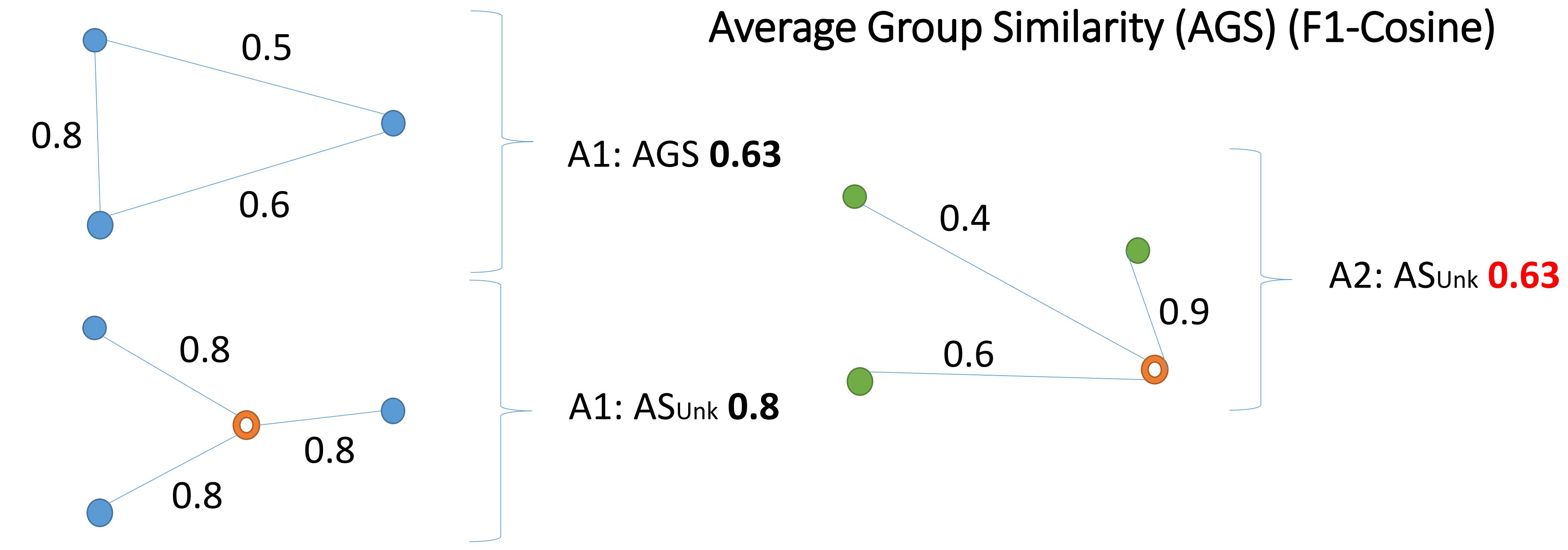
Our proposal

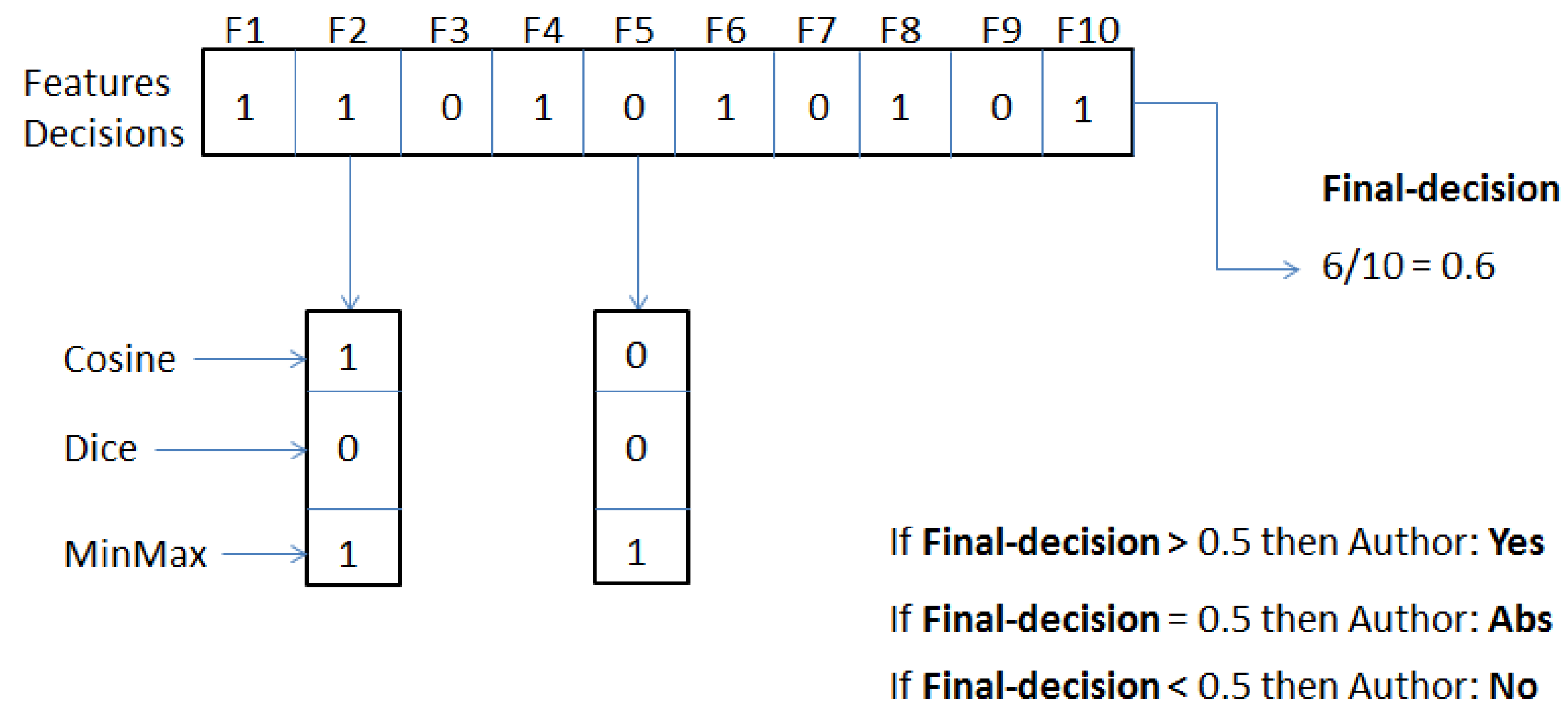
1. Representation of all documents by **one feature type**.
2. Average similarity between the document samples of an author (**AGS**).
3. Average similarity between the document of unknown authorship and the known samples of each author in a set (**AS$_{Unk}$**), in which we know who is the author that is been analyzed and the rest are used as impostors.
4. For **each linguistic feature** analyzed, we obtain a vote by majority combining the use of **different similarity functions**, in which 1 represents that the document was written by the author in analysis and 0 the opposite.
5. We obtain as a **final decision** a value in the [0, 1] interval, dividing all the votes with 1 for the features by the **total number of features** used.

Features	
Tri-grams of characters (**F1**)	_th, the, he_, …, _is, is_, _sw, swi, wim, imm, mmi, min, ing, ng.
Quad-grams of characters (**F2**)	_the, the_, …, _swi, swim, wimm, immi, mmin, ming, ing.
Word prefixes of size 2 (**F3**)	th, wh, do, is, sw
Word suffixes of size 2 (**F4**)	He, te, og, is, ng
Uni-grams of words (**F5**)	the, white, dog, is, swimming, .
Tri-grams of words (**F6**)	_ the white, the white dog, white dog is, dog is swimming,
Uni-grams of lemmas (**F7**)	the, white, dog, be, swim
Uni-grams of Part of Speech (**F8**)	DT, A, N, V, V
Tri-grams of lemmas (**F9**)	_ the white, the white dog, white dog be, dog be swim, be swim .
Tri-grams of Part of Speech (**F10**)	_ DT A, DT A N, A N V, N V V

Known sample: *The white dog is swimming.*

Unknown sample: *The white dog swim now.*





If **Final-decision** > 0.5 then Author: **Yes**
If **Final-decision** = 0.5 then Author: **Abs**
If **Final-decision** < 0.5 then Author: **No**

Results

Language	Type	Problems	Documents	Avg. known
Dutch	Cross-genre	165	452	1.74
English	Cross-topic	500	1000	1.00
Greek	Cross-topic	100	380	2.80
Spanish	Mixed	100	500	4.00

Table 1. Overview of the PAN-2015 Test Corpus

Language	ranking/ participants	User	AUC	C1	finalScore
Dutch	1/17	moreau15	0.8253	0.7697	0.63523
	13/17	castro15	0.50287	0.49091	0.24686
English	1/17	bagnall15	0.8111	0.75651	0.61361
	2/17	**castro15**	**0.74987**	**0.694**	**0.52041**
Greek	1/15	bagnall15	0.8822	0.8505	0.75031
	10/15	castro15	0.621	0.63	0.39123
Spanish	1/17	bartoli15	0.9318	0.83	0.77339
	13/17	castro15	0.5576	0.59	0.32898

Table 2. Evaluations results for authorship verification

Conclusions and future work

- We have presented the implementation of a method for authorship analysis that compares the **average similarity** calculated between a **document** of **unknown** authorship and **documents** written by an **author**, with the average similarity of the samples of this author.
- Prove as a **limit** to determine if the unknown text is of the author if his **AS$_{Unk}$ is superior to the less AS** of one of the known document sample.
- Evaluate overall **different genre of documents** if all the features or functions **contribute to the task**.

Acknowledgements

This research has been partially funded by the Spanish Ministry of Science and Innovation (TIN2012-38536-C03-03)