

INAOE at PAN 2013-2015

Our approaches for author profiling

Miguel A. Álvarez- Carmona, A. Pastor López-Monroy,
M. Montes-y-Gómez, L. Villaseñor-Pineda and H. J. Escalante

Laboratory of Language Technologies
National Institute of Astrophysics, Optics and Electronics (INAOE), MEXICO

<http://ccc.inaoep.mx/~mmontesg/>
mmontesg@inaoep.mx

The origin of our idea (PAN 2013)

- **Content** and style are important
 - It is usual to consider a great number of features
 - Some features are clearly related to some profiles (e.g., men talk more about sports, women about family)
- **Bag of features** was the common representation
 - High dimensionality and sparsity
 - Do not preserve any kind of relationship among terms.
- We proposed a **concise representation** that emphasizes the relation of terms with profiles.



A concise document representation

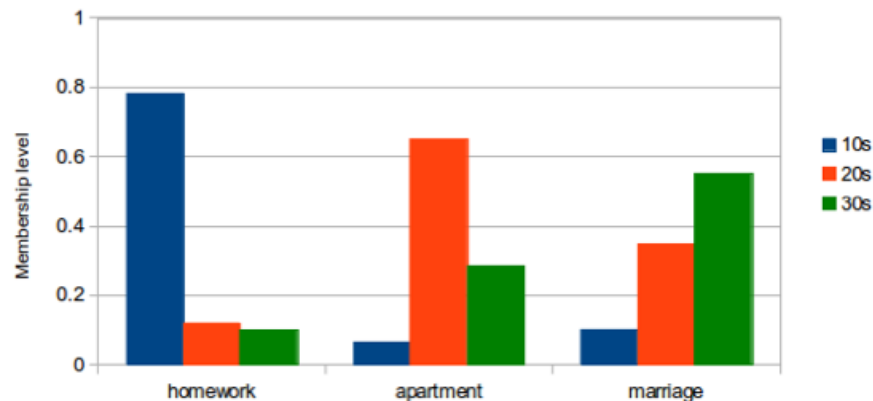
- Terms are represented by their associations profiles
- Documents' representations are the aggregation of their terms' representations

	p_1	.	.	.	p_i
t_1	$wtp_{11}(p_1, t_1)$.	.	.	$wtp_{i1}(p_i, t_1)$
.
.
.
t_j	$wtp_{1j}(p_1, t_j)$.	.	.	$wtp_{ij}(p_i, t_j)$



$$\vec{d}_k = \sum_{t_j \in D_k} \frac{tf_{kj}}{\text{len}(d_k)} \times \vec{t}_j$$

	p_1	.	.	.	p_i
d_1	$dp_{11}(p_1, d_1)$.	.	.	$dp_{i1}(p_i, d_1)$
.
.
.
d_j	$dp_{1j}(p_1, d_j)$.	.	.	$dp_{ij}(p_i, d_j)$



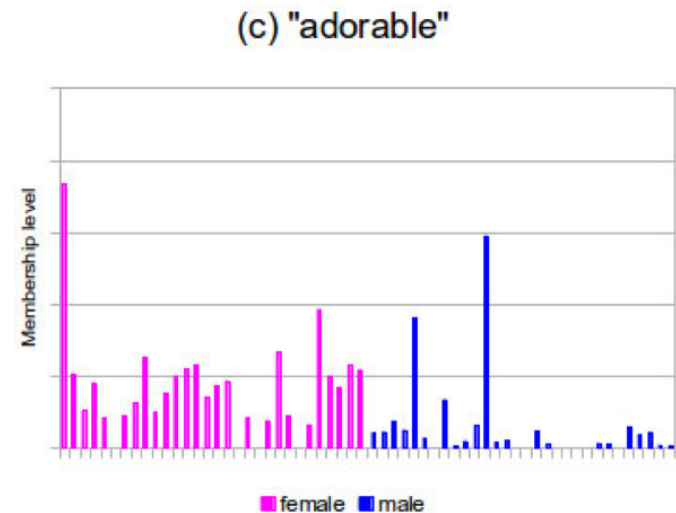
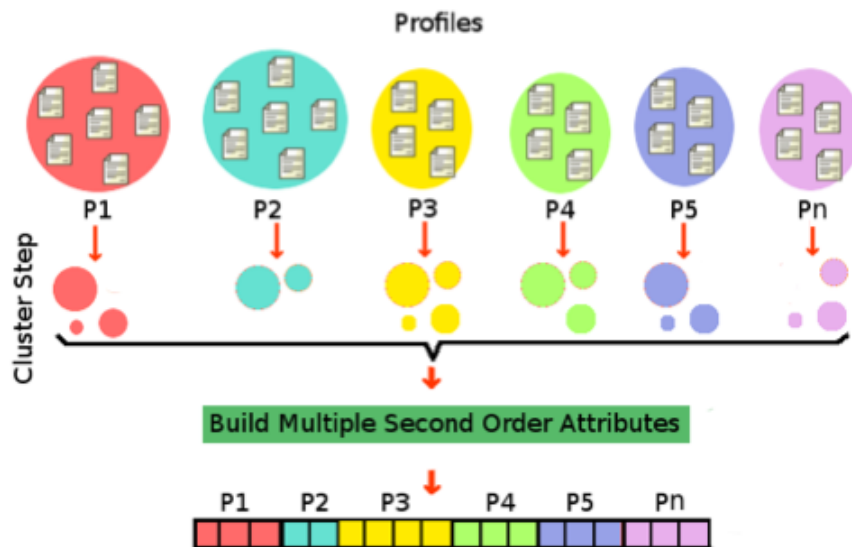
Results at PAN 2013

- **Best overall performance.**
 - English: 0.57 (gender), 0.66 (age)
 - Spanish: 0.63 (gender), 0.66 (age)
- BUT, our approach assumed certain **homogeneity** among all authors that belong to a same profile, and this is not true, especially for social media.
- Our solution for 2014: same approach but using **intra-profile information.**



Subprofile-based representation (PAN 2014)

- **Cluster** each target profile into several subprofiles.
- Build the representations of terms and documents at subprofile level.
 - As many features as the subgroups in all profiles



Results at PAN 2014

- Best overall performance
- Most important: n-SOA was better than BoT and SOA.

Age and gender prediction in the english dataset.

Dataset	Representation	Blogs		Twitter		Social Media		Reviews	
		Age	Gender	Age	Gender	Age	Gender	Age	Gender
Train	BoT	45.57	73.87	39.21	71.52	34.30	54.29	31.17	64.87
	1-SOA	46.72	75.44	43.52	70.52	35.81	55.01	32.63	66.75
	n-SOA	48.07	77.96	47.97	71.98	37.00	55.36	33.92	68.05
Test	n-SOA	39.74	67.95	49.35	72.08	35.52	52.37	33.37	68.09


Age and gender prediction in the English corpus

Dataset	Representation	Blogs		Twitter		Social Media	
		Age	Gender	Age	Gender	Age	Gender
Train	BoT	43.18	62.50	39.88	62.60	37.65	63.83
	1-SOA	45.33	62.91	41.54	62.01	38.88	64.47
	n-SOA	48.22	63.05	43.61	62.51	41.42	65.35
Test	n-SOA	48.21	58.93	53.33	60.00	45.23	64.84



Our work after PAN 2014

- We carried out an **extensive evaluation** of the proposed representations in three corpora.
- We compared n-SOA against other dimensionality reduction techniques such as **LDA** and **LSA**




ELSEVIER







Knowledge-Based Systems


Available online 2 July 2015

In Press, Corrected Proof — Note to users



Discriminative subprofile-specific representations for author profiling in social media

A. Pastor López-Monroy^a, , , Manuel Montes-y-Gómez^a, , Hugo Jair Escalante^a, , Luis Villaseñor-Pineda^a, , Efstathios Stamatatos^b, 

 [Show more](#)



New conclusions, new directions

- The proposed representations outperformed LSA and LDA (in used datasets).
- n-SOA was more than 30 times faster than LSA. Important for large scale social media applications.
- But, SOA and LSA are not highly correlated.
- They seem to capture different things:
 - SOA emphasizes discriminative features
 - LSA emphasizes descriptive features

Evaluate this at PAN 2015!



Our experiments for PAN 2015

Our participation focuses in three main goals:

- Evaluate **SOA** and **LSA** in the new collections.
- Determine if their **combination** is a good idea.
- Evaluate these representations in the classification of **personality** traits.



Results at PAN 2015

Best overall performance!

Table 3. Detailed classification accuracy to gender

Language	BOW	SOA	LSA	LSA+SOA
English	74.00	70.86	74.34	78.28
Spanish	84.00	74.00	91.00	91.00
Italian	76.31	73.68	86.84	86.84
Dutch	82.35	91.07	91.17	91.17

Table 4. Detailed classification accuracy to age

Language	BOW	SOA	LSA	LSA+SOA
English	74.83	68.21	78.94	79.60
Spanish	80.00	74.00	81.00	82.00



Results at PAN 2015

Table 5. Detailed classification accuracy for personality

Trait	English		Spanish		Italian		Dutch	
	BOW	LSA+SOA	BOW	LSA+SOA	BOW	LSA+SOA	BOW	LSA+SOA
Extroverted	64	87	62	87	65	94	64	91
Stable	56	85	69	91	52	94	61	94
Agreeable	60	80	62	84	71	92	61	88
Conscientious	61	78	62	86	57	94	67	91
Open	65	86	62	74	55	84	64	97



What we observed?

- LSA is better than SOA (we could not submit n-SOA results)
 - **Surprising** because of the collections' sizes
- Combination not relevant
 - **Surprising** because our previous results in three different collections suggest a different conclusion
- Very good results classifying personality traits
 - **Surprising** because it is a more difficult task

We decided to look at data.



Manual analysis of data

- **Very small datasets:** the 324 IDs in the corpora correspond to only **122 users**.
 - Italian: 38 IDs → 19 users
 - English: 152 IDs → 78 users
 - Spanish: 100 IDs → 49 users
 - Dutch: 34 IDs → 19 users
- Impossible to learn to distinguish men from women from only 19 examples.
 - Not a good idea because of the diversity of twitter.
- Not clear how the organizers built the training and test sets.
 - Overlap could explain the good results of BoW approach.

Split real users into several “virtual” users



Final remarks

- n-SOA seems to be a good approach for AP
 - Good results in PAN's datasets as well as in Schler et al. dataset.
- We have to continue studying the complementarity of SOA and LSA.
 - Now we do not have any strong conclusion about it.
- Recommendation: corpora has to be extended and revised, if we want to be able to obtain relevant conclusions from future editions.



Thank you for your attention!

Work in collaboration with:

Miguel A. Alvarez, Pastor López,
Hugo Escalante and Luis Villaseñor

Manuel Montes y Gómez

mmontesg@inaoep.mx

<http://ccc.inaoep.mx/~mmontesg>

