Algoritmos bio-inspirados aplicados a tareas de clasificación de textos cortos

Leticia Cagnina[1], Marcelo Errecalde[1], Paolo Rosso[2] *

[1] LIDIC (Research Group). Universidad Nacional de San Luis. Argentina.
{lcagnina,merreca}@unsl.edu.ar
[2] Natural Language Engineering Lab. - ELiRF, DSIC, Universidad Politécnica de Valencia. España. prosso@dsic.upv.es

Resumen El agrupamiento de textos cortos como así también la atribución de autoría son dos problemas típicos y de gran interés en el área de Procesamiento del Lenguaje Natural. Trabajos previos han demostrado que algoritmos bio-inspirados tales como los basados en Particle Swarm Optimization han resultado efectivos y eficientes para la resolución de problemas de agrupamiento de textos cortos. Con base en estas experiencias, se pretende aplicar este tipo de algoritmos para resolver problemas de atribución de autoría.

Palabras Clave: Agrupamiento de Textos Cortos, Atribución de Autoría, Particle Swarm Optimization.

1. Trabajo Previo: Agrupamiento de Textos Cortos

El *agrupamiento de textos cortos* es una tarea importante del Procesamiento del Lenguaje Natural ya que está presente en aplicaciones como minería de textos, extracción de información de la web, generación de textos y otras derivadas del uso de lenguajes reducidos en blogs y mensajes de textos. El objetivo de un problema de agrupamiento de textos es clasificar un conjunto de documentos en diferentes grupos. Si los documentos son cortos, el problema se dificulta debido a la baja frecuencia de términos presentes en cada texto. Este último problema es una aplicación corriente ya que permite organizar grandes volúmenes de información (expresada en textos cortos) en un número reducido de grupos significativos. En problemas de agrupamiento de textos cortos no se cuenta con información referida a los grupos ni la clasificación correcta de los documentos, dificultándose así la evaluación de una potencial solución a través de medidas externas como la *Medida F* o la *Entropía*. Como consecuencia de ello, la calidad de la solución debe ser evaluada con respecto a propiedades estructurales expresadas en medidas de validación interna como el *coeficiente de Silhouette Global*

* La investigación de la primera autora está parcialmente financiado por el programa de Estancias en la UPV de Investigadores de Prestigio PAID-02-10 N 2257. La investigación de los dos últimos autores está parcialmente financiado por el proyecto MICINN TEXT-ENTERPRISE 2.0 TIN2009-13391-C04-03 (Plan I+D+i).

(GS)[17] y la *Medida de Densidad Esperada* (DEM)[11]. Estas dos medidas son utilizadas frecuentemente ya que proveen una buena estimación de la calidad de los grupos de la solución obtenida además de aportar un buen grado de correlación con respecto a la correcta clasificación realizada por una persona. Estos dos motivos han llevado a que las medidas GS y DEM sean utilizadas como una función objetivo para optimizar por distintos algoritmos de agrupamiento.
En [10] dos versiones distintas de algoritmos basados en la técnica de optimización bio-inspirada Particle Swarm Optimization (PSO)[5] fueron presentadas para resolver el problema de agrupamiento de 3 colecciones pequeñas: *Micro4News*, *EasyAbstract* y *CICLing-2002* [6]. Una de las versiones es un algoritmo PSO discreto denominado CLUDIPSO cuya representación de soluciones es un vector de n números enteros, indicando cada uno de ellos el grupo al cual pertenece cada uno de los n documentos de la colección. La otra versión es un algoritmo PSO continuo denominado CLUCOPSO cuya representación de soluciones es un vector $(K \times T)$ de números reales donde K indica el número de centroides (uno por cada grupo de la colección) de T términos. Ambas versiones utilizan las medidas GS y EDM como función objetivo a optimizar. Las conclusiones del trabajo indican que CLUDIPSO tuvo un desempeño consistente en las 3 colecciones evaluadas logrando superar a algoritmos efectivos representativos del área como K-Means, MajorClust [18] y DBSCAN [7], cuando la medida GS fue empleada. Los mejores resultados de CLUCOPSO fueron alcanzados con la medida DEM pero siendo competitivos sólo en colecciones de mediana y alta complejidad (EasyAbstract y CICLing-2002). Los Cuadros 1, 2 y 3 ilustran los resultados obtenidos con los diferentes algoritmos. Notar que los mejores valores fueron resaltados.

Cuadro 1. Micro4News: Valores medios, mínimos y máximos de DEM y GS.

Algoritmo	DEM med	DEM min	DEM max	GS med	GS min	GS max
K-Means	0.99	0.89	1.07	0.39	0.05	0.74
MajorClust	1.08	1.05	1.10	0.69	0.64	0.74
DBSCAN	1.05	1.01	1.10	0.54	0.36	0.67
CLUDIPSO	1.07	1.06	1.07	**0.72**	**0.69**	**0.74**
CLUCOPSO	**1.11**	**1.10**	**1.12**	0.26	0.19	0.36

Cuadro 2. EasyAbstract: Valores medios, mínimos y máximos de DEM y GS.

Algoritmo	DEM med	DEM min	DEM max	GS med	GS min	GS max
K-Means	0.9	0.86	0.92	0.08	-0.05	0.29
MajorClust	0.94	**0.93**	0.96	0.31	-0.01	**0.50**
DBSCAN	0.93	0.91	0.94	0.23	0.08	0.32
CLUDIPSO	0.94	**0.93**	0.95	**0.47**	**0.44**	**0.50**
CLUCOPSO	**0.98**	0.92	**1.03**	0.25	0.21	0.32

2. Atribución de Autoría de Textos Cortos

La atribución de autoría es la tarea de determinar el autor de un texto considerando un conjunto de documentos de varios autores candidatos. Esta tarea

Cuadro 3. CICLing-2002: Valores medios, mínimos y máximos de DEM y GS.

Algoritmo	DEM med	DEM min	DEM max	GS med	GS min	GS max
K-Means	0.87	0.84	0.91	0.07	-0.06	0.22
MajorClust	0.92	0.91	0.94	0.14	-0,24	0.36
DBSCAN	0.91	0.88	0.95	0.08	-0.11	0.21
CLUDIPSO	0.92	0.91	0.93	**0.39**	**0.36**	**0.41**
CLUCOPSO	**1.07**	**1.06**	**1.11**	0.16	0.14	0.18

puede ser empleada en disputas sobre autoría de obras literarias [14], para investigaciones criminales [2] y verificación de autoría de mensajes o correos electrónicos [3], entre otras.

Una forma de resolver la tarea de atribución de autoría es utilizando las características estilográficas de escritura que posee cada autor. En [1] se identifican 3 tareas estilográficas fundamentales en aplicaciones de recuperación de la información: la caracterización del autor a través de un estilo único, la detección de similitudes y luego, la atribución de autoría. La caracterización del autor refleja información específica como género, educación, nivel social, etc. y debe ser invariante entre textos del mismo autor. La detección de similitudes se enfoca en la comparación de varios textos de forma tal de detectar propiedades comunes entre ellos. Finalmente, la atribución de autoría permitirá identificar el autor de un texto utilizando alguna medida de similitud con otros textos de autoría conocida.

Con base en las 3 tareas enunciadas, es posible automatizar la atribución de autoría mediante algoritmos que permitan: (1) representar los textos de forma tal que queden reflejadas características estilográficas propias del autor, (2) utilizar una función que permita medir similitudes entre textos de autores candidatos y el texto de autoría desconocida y, (3) decidir el autor del texto más probable.

Comúnmente, la forma más utilizada de automatizar la atribución de autoría es considerando éste como un problema de clasificación [13,8,15]. Haciendo uso del conocimiento de la buena prestación de algoritmos bio-inspirados en tareas de agrupamiento de textos cortos, actualmente se está estudiando la manera de resolver el problema de atribución de autoría, particularmente de textos cortos, utilizando estos algoritmos. El problema se transformaría en uno de agrupamiento más que de clasificación, empleando los documentos candidatos como grupos predefinidos. Luego, a través de alguna medida de similitud, el algoritmo debería ser capaz de seleccionar el grupo (uno por cada autor candidato) más probable para el documento de autoría desconocida. Se están estudiando varias formas de representación de los documentos de forma tal que se capturen las principales características estilográficas del autor. Las que se evaluarán son: uso de palabras cortas (dos o tres letras) [8], frecuencia de ocurrencia de palabras funcionales [15], utilización de $n-$gramas [9] o algunas específicas que emplean un ordenamiento de frecuencias de palabras [4]. Como función de similitud a optimizar se pretende evaluar las siguientes medidas: SVM basada en la frecuencia de las palabras [12] y la medida *Relative Hardness* [16] que utiliza el grado de

solapamiento de vocabulario entre grupos. Como conclusión de este estudio se persigue determinar si el problema de atribución de autoría puede ser abordado como uno de clasificación, con un algoritmo bio-inspirado.

Referencias

1. G. Bonanno, F. Moschella, S. Rinaudo, P. Pantano, and V. Talarico. Manual and evolutionary equalization in text mining. In *Proc. of the 7th WSEAS International Conference on Simulation, Modelling and Optimization*, pages 262–267, 2007.
2. C. Chaski. Empirical evaluations of the language-based author identification techniques. *Forensic Linguistics*, 8(1):1–65, 2001.
3. O. de Vel, A. Anderson, M. Corney, and G. Mohay. Mining e-mail content for author identification forensics. *SIGMOD Record*, 30(4):55–64, 2001.
4. L. Dinu and M. Popescu. Ordinal measures in authorship identification. In *PAN'09*, pages 62–66, 2009.
5. R. Eberhart and Y. Shi. A modified particle swarm optimizer. In *International Conference on Evolutionary Computation*. IEEE Service Center, 1998.
6. M. Errecalde and D. Ingaramo. Short-text corpora for clustering evaluation. Technical report, LIDIC, 2008.
7. M. Ester, H. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining*, pages 226–231, 1996.
8. D. I. Holmes. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguist Computing*, 13(3):111–117, 1998.
9. J. Houvardas and E. Stamatatos. N-Gram Feature Selection for Authorship Identification. volume 4183 of *LNCS*, chapter 10, pages 77–86. 2006.
10. D. Ingaramo, M. Errecalde, L. Cagnina, and P. Rosso. *Computational Intelligence and Bioengineering*, chapter Particle Swarm Optimization for Clustering short-text Corpora, pages 3–19. IOS Press, 2009. F. Masulli et al. (Eds.).
11. D. Ingaramo, D. Pinto, P. Rosso, and M. Errecalde. Evaluation of internal validity measures in short-text corpora. *Proc. of the CICLing 2008 Conference. LNCS*, 4919:555–567, 2008. Publisher Springer-Verlag.
12. M. Koppel, J. Schler, and E. Bonchek-Dokow. Measuring Differentiability: Unmasking Pseudonymous Authors. *J. of Mach. Lear. Research*, 8:1261–1276, 2007.
13. R. Matthews and T Merriam. Neural computation in stylometry i: An application to the works of shakespeare and fletcher. *Literary and Linguist Computing*, 8(4):203–209, 1993.
14. F. Mosteller and D. L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963.
15. F. Mosteller and D. L. Wallace. *Inference and Disputed Authorship: The Federalist.* Addison-Wesley, 1964.
16. D. Pinto and P. Rosso. On the relative hardness of clustering corpora. In *TSD*, pages 155–161, 2007.
17. P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987.
18. B. Stein and O.Ñiggemann. On the nature of structure and its identification. In *Proc. of the 25th International Workshop on Graph Theoretic Concepts in Computer Science - WG99, LNCS*, volume 1665, pages 122–134. Springer-Verlag, 1999.