Tópicos avanzados en categorización de textos

Marcelo Errecalde,† Diego Ingaramo,† M. Verónica Rosas,† Amparito Asensio

†Laboratorio de Investigación y Desarrollo en Inteligencia Computacional (LIDIC)[1]
Departamento de Informática
Universidad Nacional de San Luis
Ejército de los Andes 950 - Local 106
(D5700HHW) - San Luis - Argentina
Tel: (02652) 420823 / Fax: (02652) 430224
e-mail: {merreca, daingara, mvrosas, aaasensi}@unsl.edu.ar

Resumen

Contexto: Este artículo describe, en forma resumida, los trabajos de investigación y desarrollo que se están llevando a cabo en la línea "Agentes y Sistemas Inteligentes" del LIDIC, en el área de *categorización de textos*. Otras líneas de investigación del LIDIC, también abordan problemas de categorización pero, en nuestro caso, nos centramos en problemas que involucran *documentos*. Por este motivo, en nuestra línea se presta especial atención a técnicas vinculadas al procesamiento del lenguaje natural, la lingüistica computacional y la recuperación de la información. En este sentido, buena parte de los desarrollos en estos temas, se han realizado en forma conjunta con grupos de investigación con una experiencia considerable en el procesamiento del lenguaje natural, como por ejemplo, el NLEL[2] de la Universidad Politécnica de Valencia, España.

Los enfoques utilizados en nuestra línea de trabajo, buscan mejorar los procesos de categorización automática de textos en base a dos mecanismos principales: 1) el uso de técnicas de representación de textos más elaboradas, 2) el uso de algoritmos de categorización más eficientes y efectivos. Respecto al primer punto, nuestros trabajos incluyen el uso de representaciones que incorporan información semántica (conceptos) a los métodos tradicionales basados en términos y representaciones basadas en LSI (Latent Semantic Indexing). Las soluciones algorítmicas por su parte, incluyen el ensamblaje de clasificadores y los métodos de optimización bio-inspirados.

1. Introducción

Las tecnologías de la información y la comunicación han hecho disponible hoy en día, una gran cantidad de información. Gracias a Internet, es posible conseguir, sin demasiado esfuerzo, casi cualquier información que se desee en pocos segundos. Sin embargo, también es común recibir mucha más información que la que se desea (o se puede procesar), y somos inundados diariamente por una cantidad creciente de mensajes de e-mail, SMSs, reportes internos, faxes, llamadas telefónicas, noticias, artículos de revistas, páginas Web, etc [12]. Un aspecto importante a ser observado en este contexto, es que gran parte de esa información es de tipo textual, razón por la cual, todos los aspectos vinculados al procesamiento y organización automática de documentos (recuperación, categorización, agrupamiento, etc) adquieren, día a día, una relevancia creciente.

En el caso particular de la categorización automática de textos, ésta ha despertado un interés significativo, debido a la creciente disponibilidad de documentos en formato digital no sólo a través de Internet y librerías digitales sino también por el hecho de que las empresas organizan mayoritariamente sus escritos y correspondencia a partir de estos formatos. Dada la necesidad imperiosa de su organización y mantenimiento, es común hoy en día, encontrar innumerables aplicaciones de categorización de documentos que abarcan diversas áreas como la detección de "spams" [11, 22], filtrado de noticias [1], detección de plagios e identificación de autores [19, 24] análisis de opinión [2, 6], organización de patentes en categorías [18], y clasificación y organización de páginas Web [21] entre otras.

A partir del auge de esta línea de investigación,

[1]Las investigaciones realizadas en el LIDIC son financiadas por la Universidad Nacional de San Luis y por la Agencia Nacional de Promoción Científica y Tecnológica.

[2]Natural Language Engineering Lab., Departamento de Sistemas Informáticos y Computación (DSIC). Universidad Politécnica de Valencia, Valencia, España.

como una de las tareas del procesamiento del lenguaje natural, surge la necesidad de incorporar y plantear nuevas mejoras a dicho proceso, que permitan mayor exactitud al momento de seleccionar la categoría a la que pertenece un determinado texto.

Los enfoques para la construcción automática de clasificadores, pueden ser agrupados en dos grandes áreas: la categorización *supervisada* y la *no supervisada*. La categorización supervisada utiliza un proceso inductivo general basado en el conocimiento que se tiene de: a) las categorías y b) ejemplos de documentos categorizados por un experto. En la categorización no supervisada (agrupamiento) en cambio, no se conocen a priori las categorías ni asignaciones correctas de categorías (sólo se conoce alguna medida de similitud).

A la hora de realizar mejoras en cualquiera de estos procesos de categorización, los enfoques pueden ser agrupados en dos grandes categorías: a) aquellos que buscan mejorar la *representación* de los documentos, y b) los que se centran en obtener mejoras en los *algoritmos* de categorización. Los primeros, se centran en tratar de capturar en las estructuras utilizadas para representar los documentos, tanta información relevante como sea posible, que pueda ser útil para determinar la categoría de los documentos. Los segundos en cambio, intentan desarrollar algoritmos que hagan un uso eficiente de esa información, tratando de encontrar soluciones de calidad y en forma eficiente.

Actualmente, en nuestra línea de investigación se están abordando cuatro temas principales: 1) agrupamiento de documentos cortos sobre temáticas relacionadas, 2) agrupamiento de documentos cortos multilingüe, 3) categorización semántica de documentos y 4) ensamblaje de clasificadores.

En estos temas, en algunos casos se hace hincapié en los aspectos de representación y en otros casos en los aspectos algorítmicos, como se puede observar en la Tabla1. A modo de ejemplo, mientras que en la categorización semántica se ha puesto especial énfasis en enriquecer los enfoques tradicionales basados en el modelo vector (*representación*), en el agrupamiento de textos cortos los esfuerzos se han concentrado en la definición de métodos de clustering bio-inspirados (algoritmos).

El resto de este trabajo se organiza de la siguiente manera. En la sección 2, se describen brevemente los cuatro temas de investigación referidos previamente. En la sección 3 se describen algunos de los resultados obtenidos en estos temas y posibles trabajos futuros. Finalmente, la sección 4 menciona los distintos trabajos de tesis que se han desarrollado o están en ejecución, referidos a estos temas.

2. Líneas de Investigación y desarrollo

A continuación, se detallan los ejes de los temas que se están investigando, de acuerdo a los criterios especificados en la sección previa.

2.1. Agrupamiento de documentos muy cortos

El agrupamiento de documentos es la asignación de documentos a categorías desconocidas. Esta tarea es más difícil que la categorización supervisada debido a que no se dispone con anticipación, de ninguna información sobre las categorías y clasificaciones correctas de los documentos. Cuando el agrupamiento involucra textos muy cortos, la tarea es aún más dificultosa debido a las bajas frecuencias de ocurrencia de los términos en los documentos. No obstante esto, el trabajo de investigación relacionado al agrupamiento de textos cortos es muy importante, en especial si consideramos la tendencia actual/futura de la gente a usar “lenguajes pequeños”, por ejemplo en blogs, mensajes de textos, e-mails, snippets, etc. Las aplicaciones potenciales en diferentes áreas del procesamiento del lenguaje natural incluyen el re-ranking de snippets en recuperación de información y el agrupamiento automático de textos científicos disponibles en la Web.

Nosotros estamos interesados en analizar este tipo de colecciones para desarrollar técnicas novedosas que puedan ser usadas para mejorar los resultados obtenidos con técnicas de agrupamiento clásicas. Este análisis ha incluído el estudio de la correlación entre medidas de validez de clustering internas y externas, la relación de estas medidas con la dificultad inherente de las colecciones, y el desarrollo de técnicas de clustering basadas en optimización mediante enfoques bio-inspirados.

	Representación	Algoritmos
Agrupamiento textos cortos - temas relacionados		X
Agrupamiento textos cortos - multilingüe	X	X
Categorización semántica de documentos	X	
Ensamblaje de clasificadores		X

Tabla 1: Temas abordados - Aspecto enfatizado

2.2. Agrupamiento de documentos cortos multilingüe

En el clustering multilingüe, se tiene una colección con documentos en distintos lenguajes y la idea es lograr agrupamientos donde en cada grupo puede haber documentos relacionados pero de distintos lenguajes. El principal problema en este caso, es la definición de una medida de similitud entre documentos escritos en distintos lenguajes. Esto implica en general, que los documentos deban ser preprocesados para hacerlos comparables. En general existen dos formas básicas para lograr esto: 1) mediante tecnologías de traducción, 2) mediante la traducción de los documentos a una representación independiente del lenguaje.

En nuestro caso, no estamos realizando agrupamiento multilingüe estricto, sino que realizamos agrupamiento *asistido* por información multilingüe. La idea intuitiva es que si se dispone de corpus de textos paralelos en lenguaje dual (por ejemplo, en inglés y español) esta información puede servir para lograr un agrupamiento más efectivo que en el caso en que se trabaja con los documentos en cada uno de estos lenguajes por separado

2.3. Categorización semántica de documentos

La incorporación de información semántica en el indexado de documentos, es una idea que ha sido considerada en distintas áreas del Procesamiento del Lenguaje Natural, reportándose buenos resultados con este enfoque en tareas de recuperación de la información, agrupamiento y categorización de documentos. La idea principal en estos enfoques, es enriquecer las tradicionales representaciones basadas en el modelo de espacio vector, donde cada texto es representado por un vector de n-términos, siendo n el número de términos que aparecen en la colección de documentos. En los enfoques basados en información semántica en cambio, la idea es tomar en cuenta un indexado basado en los *conceptos* o *significados* de los términos que aparecen en los documentos, y reducir así los problemas introducidos por los fenómenos de *sinonimia* y *polisemia*.

Una componente fundamental en estos enfoques semánticos, es el proceso encargado de determinar en forma no ambigua, cual es el sentido (o significado) de una palabra en un contexto particular, denominado *Word Sense Disambiguation* (WSD). Si bien se han propuesto las más diversas técnicas para la tarea de WSD, podemos diferenciar en este sentido dos grandes enfoques: WSD supervisada y WSD no supervisada. Otro aspecto fundamental, es el impacto que tienen los procesos de reducción de vocabulario necesarios para reducir la dimensionalidad de los vectores usados para representar los documentos, y de que manera esta reducción se relaciona con los nuevos conceptos introducidos en la representación.

Este último aspecto es el eje principal de nuestra investigación en este tema, el cual intenta realizar aportes en el área de la categorización de textos cortos, planteando posibles mejoras a partir de la incorporación de información semántica y el uso de técnicas de reducción de vocabulario. En este sentido, se hace especial hincapié en el uso de técnicas de desambiguación no supervisadas para la obtención de los conceptos. También se analizan las distintas herramientas de la Ingeniería de Software que pueden ser útiles en el desarrollo de este tipo de aplicaciones y cual es la factibilidad de la aplicación de este tipo de técnicas en problemas complejos de la vida real, como la categorización automática de noticias.

2.4. Ensamblaje de clasificadores

La idea de considerar todas las hipótesis disponibles para resolver un problema y construir una hipótesis de más alto nivel, es un enfoque conocido en los sistemas de decisión donde es común tomar en cuenta las opiniones de varios asesores en lugar de restringirse a uno solo. Este principio ha sido aplicado en distintos métodos de minería de datos que combinan varias hipótesis en un único modelo. La idea en este caso, se basa en combinar las predicciones realizadas por las distintas hipótesis disponibles (normalmente

por votación) para clasificar ejemplos o hacer regresión sobre los ejemplos. La precisión obtenida de esta manera supera, generalmente, la precisión de cada componente individual del conjunto. A estas técnicas de combinación de hipótesis se las suele referenciar como *modelos combinados* o *métodos de ensamblaje de modelos* [5] e incluyen técnicas como *bagging* [3] y *boosting* [10].

Los sistemas de ensamblaje están fundamentados matemáticamente pero además existe una fuerte componente psicológica surgida de nuestra experiencia de la vida diaria: los usamos todo el tiempo, cada vez que recurrimos a las opiniones de otros individuos (o *expertos* en una materia) para tomar una decisión. Nuestro objetivo en este caso, es mejorar la confidencia en que estamos tomando la decisión correcta ponderando las distintas opiniones y combinándolas a través de algún proceso de pensamiento para alcanzar una decisión final.

En nuestro caso, el rol del experto es cumplido por un *clasificador*, por lo que el problema puede ser replanteado como un problema de clasificación donde cada clasificador hace una hipótesis sobre la clasificación de una instancia de datos dada, dentro de un conjunto predefinido de categorías que representan distintas decisiones. La decisión estará basada en el entrenamiento previo del clasificador usando un conjunto representativo de datos de entrenamiento para el cual las decisiones correctas son conocidas a priori.

Si bien los métodos de ensamblaje ya tienen una trayectoria considerable en tareas de clasificación general dentro de la minería de datos, su uso particular en tareas de categorización de textos es relativamente reciente [23, 20]. Este último aspecto es el eje principal de nuestro trabajo, cuyo objetivo principal es realizar una primera aproximación al problema general del uso de técnicas de ensamblaje de clasificadores en la categorización de textos. En este contexto, un objetivo parcial a cumplir será el estudio y análisis de las distintas estrategias para lograr una *diversidad* adecuada en los clasificadores individuales, un aspecto que suele ser clave para el éxito de la aplicación de estas técnicas. En este sentido, nuestro enfoque para lograr la diversidad de clasificadores estará dado por el uso de distintas codificaciones de documentos y distintas técnicas para seleccionar las palabras claves utilizadas en la codificación de los documentos.

3. Resultados esperados/obtenidos

Los principales resultados en nuestra línea de trabajo se han realizado en el agrupamiento de documentos muy cortos [14, 17, 7, 4, 15, 13] y la categorización semántica de documentos [8, 9]. En el primer caso, los trabajos futuros están orientados a extender los buenos resultados obtenidos con enfoques PSO [4, 13], a otros enfoques bio-inspirados, como por ejemplo el propuesto en [16]. En el segundo caso, se pretende realizar un estudio similar al realizado en [8, 9] pero utilizando ahora distintos enfoques de WSD no supervisada.

Para el agrupamiento de documentos cortos multilingüe, se prevee comparar enfoques basados en LSI para construir un espacio semántico multilingual, con enfoques que trabajan con los documentos separados y que utilizan medidas de similitud basadas en la correspondencia entre los documentos en el corpus paralelo. En el ensamblaje de clasificadores por su parte, los resultados preliminares han permitido observar resultados competitivos con los obtenidos por métodos clásicos como SVM, Naive Bayes y k-nn en experimentos con las colecciones de documentos más conocidas. Actualmente, nuestros experimentos se han basado en los algoritmos AdaBoost y Bagging, proyectándose en el futuro hacer uso de enfoques basados en la teoría de Dempster-Shafer.

4. Formación de recursos humanos

Trabajos de tesis vinculados con las temáticas descriptas previamente:

- 1 tesis doctoral en ejecución (co-dirección con investigador del NLEL (UPV))
- 1 tesis de maestría en ejecución (co-dirección con investigador del NLEL (UPV))
- 1 tesis de Licenciatura aprobada.
- 1 tesis de Licenciatura en ejecución.

Referencias

[1] G. Amati, D. DÁloisi, V. Giannini, and F. Ubaldini. A framework for filtering news and managing distributed data. *Journal of Universal Computer Science*, pages 1007–1021, 1997.

[2] A. Balahur and A. Montoyo. Determining the semantic orientation of opinions on products- a comparative analysis. *Procesamiento del Lenguaje Natural*, pages 201–208, 2008.

[3] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–146, 1996.

[4] L. Cagnina, M. Errecalde, D. Ingaramo, and P. Rosso. A discrete particle swarm optimizer for clustering short-text corpora. In *Proc. of the 3rd International Conference on Bioinspired Optimization Methods and their Applications (BIOMA08)*, pages 93–103. 2008.

[5] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40(2):139–157, 2000.

[6] X. Ding, B. Liu, and P. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the international conference on Web search and web data mining*, pages 231–240, Palo Alto, California, USA, 2008.

[7] M. Errecalde, D. Ingaramo, and P. Rosso. Proximity estimation and hardness of short-text corpora. In *Proc. 5th International Workshop on Text-based Information Retrieval (TIR-2008)-Dexa 2008*, pages 15–19. IEEE Computer Society, 2008.

[8] E. Ferretti, M. Errecalde, and P. Rosso. The influence of semantics in text categorisation: A comparative study using the k nearest neighbours method. In *Proc. of the 2nd Indian International Conference on Artificial Intelligence (IICAI)*, 2005.

[9] E. Ferretti, M. Errecalde, and P. Rosso. Does semantic information help in the text categorisation task? *Journal of Intelligent Systems*, 17(1–3), 2008.

[10] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the 13th International Conference on Machine Learning, ICML'96*, pages 148–156, 1996.

[11] D. Heckerman, M. Sahami, S. Dumais, and E. Horvitz. A bayesian approach to filtering junk e-mail. In *Proceeding of AAAI-98 Workshop on Learning for Text Categorization*, pages 55–62, Madison, Wisconsin, USA, 1998.

[12] F. Heylighen. Information overload, complexity and information overload in society: Why increasing efficiency leads to decreasing control. http://pespmc1.vub.ac.be/papers/info-overload.pdf, CLEA, Free University of Brussels, Pleinlaan, 2002.

[13] D. Ingaramo, M. Errecalde, L. Cagnina, and P. Rosso. *Computational Intelligence and Bioengineering*, chapter Particle Swarm Optimization for clustering short-text corpora. KBIES. IOS press, 2009.

[14] D. Ingaramo, M. Errecalde, and P. Rosso. Medidas internas y externas en el agrupamiento de resúmenes cientícos de dominios reducidos. *Procesamiento del Lenguaje Natural*, 39, 2007.

[15] D. Ingaramo, M. Errecalde, and P. Rosso. Density-based clustering of short-text corpora. *Procesamiento del Lenguaje Natural*, 41:81–87, 2008.

[16] D. Ingaramo, G. Leguizamón, and M. Errecalde. Adaptive clustering with artificial ants. *Journal of Computer Science & Technology*, 5(4):264–271, 2005.

[17] D. Ingaramo, D. Pinto, P. Rosso, and M. Errecalde. Evaluation of internal validity measures in short-text corpora. *LNCS*, 4919:555–567, 2008.

[18] L. Larkey. A patent search and classification system. In *Proceedings of DL-99, 4th ACM Conference on Digital Libraries*, pages 179–187, Berkeley, CA, USA, 1999.

[19] R. Lukashenko, V. Graudina, and V. Grundspenkis. Computer-based plagiarism detection methods and tools: an overview. In ACM, editor, *Proceedings of the international conference on Computer systems and technologies, CompSysTech '07*, pages 1–6, New York, NY, USA, 2007.

[20] A. Montejo and L. A. Ureña. Binary classifiers versus adaboost for labeling of digital documents. In *Procesamiento del Lenguaje Natural*, pages 319–326, 2006.

[21] H. Oh, S.H. Myaeng, and M.H. Lee. A practical hypertext categorization method using links and incrementally available class information. In *Proceedings of SIGIR-00, 23rd ACM International Conference on Research and Development in Information Retrieval*, pages 264–271, Athens, Greece, 2000.

[22] G. Sakkis and I. Androutsopoulos. A memory-based approach to anti-spam filtering for mailing lists. *Information Retrieval*, pages 49–73, 2003.

[23] Robert E. Schapire and Yoram Singer. BoosTexter: A boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.

[24] B. Stein, M. Koppel, and E. Stamatatos. Plagiarism analysis, authorship identification, and near-duplicate detection. *SIGIR Forum*, 2007.