Minería de Textos y de la Web

Leticia Cagnina,* Edgardo Ferretti, M. Paula Villegas, M. José Garciarena, Sergio Burdisso,** Darío Funez, Carlos Velázquez, Marcelo Errecalde

Laboratorio de Investigación y Desarrollo en Inteligencia Computacional
Departamento de Informática, Universidad Nacional de San Luis
Ejército de los Andes 950 - (D5700HHW) San Luis - Argentina
e-mails de contacto: {lcagnina, ferretti, merreca}@unsl.edu.ar

Resumen

Este artículo describe, brevemente, las tareas de investigación y desarrollo que se están llevando a cabo en la línea de investigación "Minería de Textos y de la Web" en el marco del proyecto "Aprendizaje automático y toma de decisiones en sistemas inteligentes para la Web". La línea aborda diversas áreas vinculadas a la ingeniería del lenguaje natural, como por ejemplo el Procesamiento del Lenguaje Natural (PLN), la Lingüística Computacional, la Minería de Textos, la Minería de la Web y la recuperación de información de la Web. En el contexto de este proyecto por lo tanto, esta línea se centra en todos los problemas vinculados con el desarrollo de herramientas inteligentes para la extracción, análisis y validación de contenido Web, que incluyen: representación de documentos y usuarios de la Web, medidas de calidad de información para el contenido Web, técnicas abiertas de extracción de información para la Web, algoritmos de categorización supervisados, semi-supervisados y no supervisados y caracterización de usuarios, entre otros.

Palabras clave: Minería de Textos, Minería de la Web, Lingüística Computacional, Procesamiento del Lenguaje Natural

*Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
**Becario de CONICET

Contexto

La línea de investigación "Minería de Textos y de la Web" es una de las tres líneas del proyecto titulado "Aprendizaje automático y toma de decisiones en sistemas inteligentes para la Web", un nuevo proyecto que será presentado este año como continuación del Proyecto de Investigación Consolidado (PROICO) titulado "Herramientas y mecanismos para la toma de decisiones en agentes inteligentes artificiales". Este último proyecto, aprobado por evaluadores externos a la UNSL, se desarrolla en el *Laboratorio de Investigación y Desarrollo en Inteligencia Computacional* (LIDIC) de la UNSL y ha sido financiado en forma directa por la UNSL (PROICO 30312) y en forma indirecta por: a) el Programa de Incentivos (22/F237), b) la Comisión Europea de Investigación e Innovación, a través del programa Marie Curie Actions: FP7 People 2010 IRSES, c) el CONICET, a través de un investigador asistente y becas: dos de Doctorado y dos de Post-Doctorado asignadas a integrantes del proyecto y d) el Consejo Nacional de Ciencia y Tecnología (CONACYT-México) y otros organismos científicos del gobierno mexicano, en los que distintos integrantes han participado en tres proyectos de investigación como colaboradores externos.

Este proyecto posee además dos líneas de investigación denominadas: "Aplicaciones" y "Toma de decisiones y aprendizaje automá-

tico"; la primera enfocada en el uso del aprendizaje automático en psicología, educación y el cuidado de la salud y la segunda dedicada al desarrollo de modelos formales y mecanismos para la toma de decisiones y aprendizaje en agentes artificiales inteligentes. Es claro en este contexto, que muchos problemas y aplicaciones intersectan los alcances de más de una de las líneas de este proyecto, lo cual involucra un trabajo integrado y coordinado permanente a los fines de optimizar los recursos disponibles para la obtención de los objetivos propuestos.

Introducción

En la actualidad, la cantidad de información disponible en la Web crece a un ritmo exponencial. Mucha de esta información está almacenada en forma de documentos de texto generados por diferentes usuarios los cuales poseen diversas características. Ejemplos de este tipo de material es el producido en las redes sociales como Facebook, Google+, sitios de microblogging como Twitter y las innumerables facilidades de chats disponibles hoy en día. La posibilidad de analizar toda la información disponible significa un reto muy importante para los investigadores de las Ciencias Sociales, razón por la que surge la necesidad de contar con herramientas automáticas que permitan acceder, organizar y almacenar el caudal de material con el que se cuenta. En este sentido, la *caracterización del autor* (en inglés, *author profiling*) es la tarea que tiene como principal objetivo el análisis de los textos de un autor con la finalidad de obtener tanta información como sea posible respecto de la/s persona/s que escribieron dichos textos. Información relacionada a la edad, género, personalidad, demografía, idioma natal y antecedentes culturales [7], son algunos ejemplos del tipo de información que se puede extraer considerando sólo los textos de una persona. El determinar correctamente el perfil de un autor es un problema que tiene una amplia gama de aplicaciones que podrían impactar en nuestras vidas de forma considerable. Por ejemplo, en marketing, el detectar características específicas de los usuarios (por ejemplo el género de personas que gustan de cierto producto), permitiría mostrar sólo ciertos tipos de productos a sólo ciertos tipos de usuarios. De manera similar, en el área de Business Intelligence, conocer qué tipo de personas son las interesadas en determinados servicios podría significar el éxito o fracaso de la empresa. Más interesante aún, en el campo forense, el reconocimiento del perfil lingüístico de acosadores (por ejemplo pedófilos) puede significar el hecho de identificar e incluso sentenciar a los sospechosos [20, 12].
Wikipedia, la enciclopedia en línea de libre acceso más popular e importante de todos los tiempos, es otra de las fuentes fundamentales de información en la Web. Sin embargo, su popularidad también conlleva un reto crucial: mejorar continua y sistemáticamente la calidad de los textos que la componen. Este aspecto no es casual si consideramos que los autores que contribuyen con Wikipedia son heterogéneos, en cuanto al nivel de educación, edad, cultura, habilidades del lenguaje y especialización en un área. De allí la importancia de poder identificar de manera automática ciertos aspectos de calidad como por ejemplo: la exactitud, fiabilidad y relevancia [21] de la información publicada. Diferentes herramientas han sido propuestas para la clasificación de los documentos de Wikipedia, como así también diferentes métricas para evaluar la presencia o no de diferentes fallas de calidad.

En la siguiente sección, se describen los principales enfoques desarrollados por los integrantes del grupo en lo que respecta al estudio de la caracterización del autor, calidad de información en Wikipedia y algunas extensiones de trabajos ya desarrollados como la clasificación no supervisada de textos cortos.

Desarrollo e Innovación

En términos generales, la línea "Minería de Textos y de la Web" se desarrolló siguiendo tres aristas bien marcadas. Cada una de ellas se detalla brevemente a continuación.

Mecanismos automáticos para la caracterización del autor (CA)
La CA basada exclusivamente en las características presentes en el texto que una persona ha escrito, ha sido una tarea muy interesante de llevar a cabo. Se han obtenido buenos resultados con técnicas estilográficas como los n-gramas de caracteres o algunas más avanzadas de segundo orden [23, 16] para la representación de los documentos. También el uso de *perfiles* con las características más importantes de cada grupo etario ha arrojado buenos resultados [15]. Actualmente, se trabaja en la búsqueda de nuevas estrategias de representación que consideren al *usuario* en un contexto más general que como el simple autor de un documento. La idea en este caso es considerar e integrar toda aquella información disponible que surge de la interacción del usuario con los medios sociales. No sólo consideraremos atributos léxicos, estilométricos o socio-lingüísticos presentes en los documentos, sino también atributos *multi-modales* como los derivados del grafo de contactos de un usuario, imágenes y videos que comparte en la Web, etc.

Calidad de la información en la web
Debido al fácil acceso a la información que existe en la actualidad a través de diferentes recursos, la evaluación de la calidad de la información en la Web se ha convertido en una tarea muy importante. Día a día tanto las personas comunes como empresas y entidades gubernamentales o privadas toman decisiones basándose en la información disponible en la Web. Esto, sumado al notable incremento de información disponible en Internet ha provocado una necesidad imperiosa de evaluar la calidad de dicha información de forma automática.

En este sentido se ha trabajado en la identificación y definición de diferentes aspectos relacionados a la calidad de información del contenido Web como confiabilidad, objetividad, especificidad, etc. Para ello se utilizaron como referencia diferentes propuestas existentes para el área de calidad de información en la Web [24, 18, 8]. Se desarrollaron características (*features*) basadas en información factual [17] y variantes del algoritmo PU-learning [19] que obtuvieron resultados muy interesantes en la clasificación de fallas de calidad en el contexto de Wikipedia [14, 13]. PU-learning es un algoritmo perteneciente al paradigma de aprendizaje semi-supervisado, ya que utiliza archivos no etiquetados para ayudar al clasificador en la distinción de la clase positiva. Los enfoques de clasificación one-class [22] también pertenecen a este paradigma de aprendizaje y en particular, la predicción de fallas de calidad en Wikipedia ha sido caracterizada como un problema one-class [1, 6, 4, 5, 2], por el grupo de investigación alemán[1] que dio origen a una línea de investigación que lleva el mismo nombre. De acuerdo con los reportes realizados por Anderka et al. [1, 2] existen diez fallas de calidad que comprenden aproximadamente el 75 % de documentos en Wikipedia, y es por eso que su predicción ha sido motivo de investigaciones recientes, principalmente a partir de la primera Competencia Internacional de Predicción de Fallas de Calidad en Wikipedia [3], realizada en el año 2012, en la que nuestro grupo obtuvo los mejores resultados.

Actualmente, estamos extendiendo medidas de calidad basadas en información factual, de manera tal de detectar fallas de calidad específicas. En este contexto, se están realizando pruebas con el subconjunto de fallas de calidad de Wikipedia en inglés denominado *Original Research* (una de las diez fallas más importantes, mencionadas precedentemente) para determinar la efectividad de este tipo de features.

[1] `https://www.uni-weimar.de/en/media/chairs/webis/home/`

Categorización no supervisada
Con el objetivo de extender los trabajos previos de algunos de los integrantes del grupo en relación al clustering de textos cortos [10] a textos más generales, se establecieron dos líneas de trabajo. En primer lugar, se extendieron estos trabajos a documentos de longitud arbitraria [11]. Luego, se buscó determinar el grado de escalabilidad de los métodos ya desarrollados, y en particular aquellos basados en enfoques de Inteligencia Colectiva [9]. En este mismo contexto, estamos analizando implementaciones más eficientes de algoritmos como *Sil-Att* [11], mediante modificaciones de la implementación del Coeficiente de Silueta y una versión adaptativa de este mismo coeficiente.

Formación de Recursos Humanos

Trabajos de tesis vinculados con las temáticas descritas previamente:

- 2 tesis de Licenciatura defendidas en 2015.
- 2 tesis de Licenciatura a iniciarse en Marzo de 2016.
- 2 tesis de Maestría en ejecución.
- 1 tesis de Doctorado en ejecución con una beca de CONICET.

Referencias

[1] M. Anderka. *Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia.* Dissertation, Bauhaus-Universität Weimar, 2013.

[2] M. Anderka and B. Stein. A Breakdown of Quality Flaws in Wikipedia. In C. Castillo, Z. Gyongyi, A. Jatowt, and K. Tanaka, editors, *2nd Joint WICOW/AIRWeb Workshop on Web Quality*, pages 11–18. ACM, 2012.

[3] M. Anderka and B. Stein. Overview of the 1st International Competition on Quality Flaw Prediction in Wikipedia. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *Working Notes Papers of the CLEF 2012 Evaluation Labs*, 2012.

[4] M. Anderka, B. Stein, and N. Lipka. Detection of Text Quality Flaws as a One-class Classification Problem. In B. Berendt, A. de Vries, W. Fan, C. Macdonald, I. Ounis, and I. Ruthven, editors, *20th ACM International Conference on Information and Knowledge Management*, pages 2313–2316. ACM, 2011.

[5] M. Anderka, B. Stein, and N. Lipka. Towards Automatic Quality Assurance in Wikipedia. In S. Srinivasan, K. Ramamritham, A. Kumar, M. Ravindra, E. Bertino, and R. Kumar, editors, *20th International Conference on World Wide Web*, pages 5–6. ACM, 2011.

[6] M. Anderka, B. Stein, and N. Lipka. Predicting Quality Flaws in User-generated Content: The Case of Wikipedia. In B. Hersh, J. Callan, Y. Maarek, and M. Sanderson, editors, *35th International ACM Conference on Research and Development in Information Retrieval*, pages 981–990. ACM, 2012.

[7] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler. Automatically profiling the author of an anonymous text. *Commun. ACM*, 52(2):119–123, 2009.

[8] R. Baeza-Yates. User generated content: how good is it? In *3rd Workshop on information credibility on the Web*, 2009.

[9] L. C. Cagnina, M. Errecalde, D. Ingaramo, and P. Rosso. An efficient particle swarm optimization approach to cluster short texts. *Information Science*, 265:36–49, 2014.

[10] M. Errecalde, D. Ingaramo, and P. Rosso. ITSA*: An effective iterative method for short-text clustering tasks. In *Proc. of the 23rd International Conference on Industrial Engineering and other Applications of Applied Intelligent Systems, IEA/AIE 2010*, volume 6096 of *LNCS*, pages 550–559. Springer-Verlag, 2010.

[11] M. L. Errecalde, L. C. Cagnina, and P. Rosso. Silhouette + attraction: A simple and effective method for text clustering. *Natural Language Engineering*, FirstView:1–40, 2 2016.

[12] H. J. Escalante, E. Villatoro-Tello, A. Juarez, M. M. y Gomez, and L. Villasenor. Sexual predator detection in chats with chained classifiers. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 46–54. ACL, 2013.

[13] E. Ferretti, M. L. Errecalde, M. Anderka, and B. Stein. On the use of reliable-negatives selection strategies in the pu learning approach for quality flaws prediction in wikipedia. In H. Decker, L. Lhotská, S. Link, M. Spies, and R. R. Wagner, editors, *DEXA Workshops*, pages 211–215. Springer, 2014.

[14] E. Ferretti, D. H. Fusilier, R. Guzmán-Cabrera, M. M. y Gómez, M. Errecalde, and P. Rosso. On the use of pu learning for quality flaw prediction in wikipedia. In P. Forner, J. Karlgren, and C. Womser-Hacker, editors, *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[15] D. G. Funez, L. Cagnina, and M. L. Errecalde. Determinación de género y edad en blogs en español mediante enfoques basados en perfil. In *XVIII Congreso Argentino de Ciencias de la Computación*, pages 984–993, 2013.

[16] M. J. Garciarena Ucelay, M. P. Villegas, L. C. Cagnina, and M. L. Errecalde. Cross domain author profiling task in spanish language: An experimental study. *Journal of Computer Science and Technology*, 41(2):122–128, 2015.

[17] E. Lex, M. Voelske, M. Errecalde, E. Ferretti, L. Cagnina, C. Horn, B. Stein, and M. Granitzer. Measuring the quality of web content using factual information. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*. ACM, 2012.

[18] A. Lih. Wikipedia as participatory journalism: reliable sources? metrics for evaluating collaborative media as a news resource. In *5th international symposium on online journalism*, 2004.

[19] B. Liu, Y. Dai, X. Li, W. Lee, and P. Yu. Building text classifiers using positive and unlabeled examples. In *3rd IEEE international conference on data mining*. IEEE Computer Society, 2003.

[20] I. McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. McBride, and E. Jakubowski. Learning to identify internet sexual predation. *International Journal of Electronic Commerce*, 15(3):103–122, 2011.

[21] T. Redman. *Data Quality for the Information Age*. Artech House, 1996.

[22] D. Tax. *One-class classification*. Ph.d. thesis, Delft University of Technology, 2001.

[23] M. P. Villegas, M. J. Garciarena Ucelay, M. L. Errecalde, and L. Cagnina. A spanish text corpus for the author profiling task. In *XX Congreso Argentino de Ciencias de la Computación*, pages 621–630, 2014.

[24] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 1996.