Sistemas Inteligentes para la Evaluación de la Calidad de la Información en la Web

Marcelo Errecalde, Edgardo Ferretti, Diego Ingaramo, María Rosas, Leticia Cagnina, Dario Funez, Patricia Roggero, Carlos Velázquez

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:{merreca,ferretti,daingara,mvrosas}@unsl.edu.ar,
{lcagnina,dgfunez,proggero}@unsl.edu.ar, carvear20@yahoo.com.ar

Resumen

En este artículo se describen, en forma resumida, los trabajos de investigación y desarrollo que se están llevando a cabo en la línea de investigación "Sistemas Inteligentes" en las áreas de *Text Mining*, *Web Mining* y *Web Intelligence*, y que abordan principalmente tareas como: clustering de textos cortos multiligües, categorización semántica de textos, medidas de calidad de la información en la Web, detección de plagio y atribución de autoría, minería de opinión y sentimientos, integración de agentes y técnicas de minería de textos, y uso de arquitecturas cognitivas para agentes web; en especial aquellas basadas en lógica, razonamiento argumentativo y teoría de decisión cualitativa. En particular, pondremos especial énfasis en aquellas problemáticas que se están comenzando a investigar en forma conjunta con investigadores de Alemania, Austria, España y Grecia en el contexto de un proyecto FP7, recientemente aprobado en la Unión Europea.

Palabras clave: Text y Web Mining, Web Intelligence, Sistemas Inteligentes, Agentes Autónomos.

Contexto

La línea de investigación "Sistemas Inteligentes" forma parte del proyecto "Nuevas tecnologías para el tratamiento integral de datos multimedia", Proyecto de Investigación consolidado de la Universidad Nacional de San Luis, que se centra en la incorporación de información no estructurada (texto, audio, imágenes y video) en la resolución de problemas y la toma de decisiones. Este proyecto recibe financiamiento de la Universidad Nacional de San Luis y de la Comisión Europea de Investigación e Innovación (Marie Curie Actions: FP7-People-2010-IRSES).

Introducción

Hoy en día, los bancos de datos y la información disponible en la Web enfatizan la cantidad de información en vez de su calidad. Este hecho es de fácil comprobación si se tiene en cuenta el aumento constante de blogs, el crecimiento notable de datos creados artificialmente, el síndrome establecido de copiar y pegar, y la falta de datos semánticamente enriquecidos. Asimismo, el uso indebido de información (ya sea intencional o no intencional), como por ejemplo actividades

de vandalismo en Wikipedia, los spam blogs (Splogs), el plagio, etc., influye significativamente en la disminución de la calidad de la información en la Web. Así, esta información descentralizada y de baja calidad conduce a problemas como:

- La búsqueda de información requiere de métodos robustos para la eliminación de información de baja calidad que no es creíble.
- Juzgar la calidad, credibilidad y fiabilidad de la información sigue siendo una tarea manual muy ardua.
- Los usuarios difícilmente pueden estimar la credibilidad de personas virtuales para establecer con ellos relaciones de confianza.
- Separar hechos contradictorios o información desactualizada de información valiosa, se convierte en un gran desafío para los sistemas de información.
- La información se almacena de forma redundante y muy dispersa en diferentes lugares.

En general, se puede afirmar que hoy en día la Web carece de mecanismos de filtrado de calidad de información, de identificación automática de uso indebido de patrones, como así también de herramientas para establecer la confianza del usuario en la información y sus autores. Es por eso que resulta de gran interés desarrollar medidas de calidad de información, métodos de detección de plagio y atribución de autoría, y de minería de opinión y sentimientos.

Desarrollo de Medidas de Calidad de la Información en la Web

La evaluación de la calidad de la información es una tarea importante, porque las decisiones a menudo se basan en información procedente de múltiples fuentes y a veces desconocidas, aunque la fiabilidad y exactitud de la información sea cuestionable.

En la literatura, se ha reportado un gran número de aspectos de calidad de información [19], y una interpretación ampliamente aceptada de la calidad de información, es que en sí mismo, es un concepto multi-dimensional que se define por ciertos aspectos de calidad (dimensiones); como por ejemplo: la exactitud, fiabilidad y relevancia [13].

Actualmente, la producción científica en lo que respecta a la investigación en calidad de la información se ha centrado en la evaluación de la calidad en datos estructurados, como por ejemplo, bases de datos relacionales y almacenes de datos, e incluye la integración de datos, verificar su integridad y la depuración de datos y gestión de transacciones [1, 9]. Sin embargo, en los últimos años la cantidad de datos no-estructurados o semi-estructurados se ha incrementado notablemente. Por lo tanto, los retos actuales para este campo de investigación reside en el desarrollo de técnicas para medir la calidad de la información para este tipo de datos.

Por esta razón, nuestro grupo de investigación en conjunto con otros grupos con objetivos afines,[1] han comenzado a trabajar en el desarrollo de algoritmos y métodos inteligentes para medir la calidad de la información en contenidos web semi-estructurados o totalmente desestructurados. De esta interacción, se ha concluido que antes de abordar el desarrollo propiamente dicho de tales métodos y algoritmos, es conveniente distinguir aspectos de calidad y sus respectivas medidas, como ser la confiabilidad de la información, su objetividad, accesibilidad, precisión, consistencia (con otras fuentes), credibilidad, integridad y la reputación o confianza.

Detección de Plagio y Atribución de Autoría

Debido a la facilidad con que se puede encontrar y manipular los textos en la Web,

[1] Know-Center Graz (Austria), Bauhaus University Weimar (Alemania), Universidad Politécnica de Valencia (España) y University of the Aegean (Grecia).

el plagio se ha incrementado notablemente. El plagio atenta contra los principios sobre los que se fundamenta el aprendizaje y la enseñanza, y representa una falta común. Sin embargo, el problema cruza la frontera de las aulas y el impacto de la Web. Portales y blogs a menudo plagian el contenido de otros sitios. La detección automática de plagio tiene como objetivo generar la tecnología que ayude a los expertos (profesores, administradores de sitios Web y lingüistas forenses, entre otros) de manera eficaz y fiable a identificar los casos de plagio en textos no-estructurados. Sin embargo, los casos de plagio pueden ser difíciles de detectar debido a la paráfrasis plagiaria y el resumen del texto de origen, como así también el uso de fuentes multilingües.

Las herramientas computacionales para ayudar en la detección de plagio se pueden agrupar en externas o intrínsecas. En la detección de plagio externa de un texto sospechoso, al mismo se lo compara con un conjunto grande de documentos fuentes potenciales con el fin de descubrir posibles plagios [18]. Para la detección de plagio externo se han propuesto diversos enfoques, desde la detección de copias exactas [10] hasta la detección de copias modificadas [17]. En lo que respecta a copias modificadas, se debe prestar especial atención a la detección de plagio multilingüe, donde el pasaje de texto sospechoso y la colección de referencia se encuentran en diferentes idiomas [12]. Este tipo de plagio ocurre cada vez más a menudo, debido a la gran cantidad de documentos que se publican actualmente en inglés y hablantes de lenguas menos difundidas toman esos recursos como base para su texto.

En la detección de plagio intrínseca se supone que ninguna colección de documentos fuente está disponible. Esto es realista, porque a menudo es difícil recuperar la fuente de un posible caso de plagio [16]. En este caso, las características estilométricas de la autoría se consideran con el fin de detectar los fragmentos de un documento sospechoso que podría ser plagiado [20].

Minería de Opinión y Sentimientos

El desarrollo de repositorios de información en línea, tales como weblogs o blogs es una de las principales razones por la cual la antigua World Wide Web ha evolucionado en lo que se conoce actualmente como la Web 2.0. Este hecho crea muchas oportunidades y también problemas para recuperar la información correcta en el momento adecuado. Uno de estos problemas se refiere al análisis de la información *subjetiva*, como por ejemplo *opiniones*. Es bien sabido que la web es actualmente la principal plataforma para la búsqueda de información. El contenido generado por el usuario (audio, imágenes, texto) constituye conocimiento pertinente y actualizado para la toma de decisiones, ya que refleja en líneas generales las tendencias actuales, patrones de comportamiento, preferencias de los usuarios, etc. Sin embargo, la información publicada por los usuarios en la mayoría de los casos se basa en puntos de vista subjetivos. Por lo tanto, la importancia de tomar en consideración este tipo de información es relevante para la minería de información cualitativa de la web.

La minería de opinión y el análisis de sentimientos trata con el análisis de opiniones con el fin de recuperar conocimientos que provienen de fuentes tales como blogs y foros. De acuerdo con [11], la primera se centra en la extracción y análisis de sentencias sobre diversos aspectos de items dados, mientras que el segundo hace énfasis en la clasificación de las opiniones de acuerdo a su polaridad (positiva o negativa).

Líneas de Investigación y Desarrollo

Los métodos comunes para evaluar la calidad de la información se basan en cuestionarios, que son contestados de forma manual por consumidores humanos para estimar aspectos de calidad de la información [2, 8].

La evaluación manual es muy larga y costosa, y debido a la enorme cantidad de información disponible sólo pueden evaluarse muestras parciales.

De esta manera, una de las temáticas abordadas por la línea de "Sistemas Inteligentes" tiene por objeto el desarrollo de algoritmos para calcular automáticamente los aspectos de calidad de información, que resulta mucho más barato que la evaluación manual y que además puede hacer frente a grandes cantidades de datos. Para lograr esta meta se piensan seguir los siguientes objetivos parciales:

- Desarrollar medidas de calidad de la información en la Web y analizar métricas de combinación de las mismas para lograr una calificación única.
- Desarrollar métodos escalables para la detección de plagio y atribución de autoría con fuentes multilingües, con el fin de medir la confiabilidad de la información en la web y medir los efectos del síndrome de copiar y pegar.
- Desarrollar métodos escalables para realizar minería de opiniones y análisis de sentimientos con fuentes multilingües, para tratar de determinar la objetividad de contenidos textuales en la Web.

Resultados y Objetivos

En este contexto, la línea actualmente se está dedicando a llevar a cabo las siguientes tareas:

1. Estudio, análisis, diseño e implementación de técnicas de aprendizaje de máquina para problemas de minería de textos. A tal fin se han utilizado métodos bio-inspirados para realizar tareas de clustering de textos cortos en inglés y con fuentes multilingës, como así también se han usado enfoques estadísticos para realizar tareas de categorización semántica de textos.
2. Estudio, análisis, diseño e implementación de agentes inteligentes para resolver aplicaciones de Web Mining y Web Intelligence. Algunos esquemas de desarrollo considerados consisten en dotar a los agentes Web con arquitecturas cognitivas, integrar estos agentes con técnicas de minería de textos, y con modelos formales de toma de decisiones basadas en argumentación.

Con respecto al primer punto, uno de los principales resultados del grupo ha consistido en el diseño e implementación de varios algoritmos bio-inspirados, que han obtenido lo mejores resultados reportados en clustering de textos cortos [6, 4, 7, 3, 5]. Asimismo, para el segundo punto, en [14, 15] se desarrolló la primera arquitectura concreta de agente que integra al modelo BDI con servicios Web y razonamiento argumentativo en un mismo framework.

Formación de Recursos Humanos

Trabajos de tesis vinculados con las temáticas descritas previamente:

- 1 tesis Doctoral en ejecución en co-dirección con investigador de la Universidad Politécnica de Valencia (UPV).
- 2 tesis de Maestría en ejecución (una en co-dirección con investigador de la UPV).
- 1 tesis de Licenciatura aprobada.

Referencias

[1] D. P. Ballou, I. N. Chengalur-Smith, and R. Y. Wang. Sample-based quality estimation of query results in relational database environments. *IEEE Transactions on Knowledge and Data Engineering*, 18(5):639–650, 2006.

[2] M. Bobrowski, M. Marre, and D. Yankelevich. A homogeneous framework to measure data quality. In *Proc. of the International Conference on Information Quality*, 1999.

[3] L. Cagnina, M. Errecalde M., and P. Rosso. Algoritmos bio-inspirados aplicados a tareas de clasificación de textos cortos. In *IV Jornadas TIMM (Temática en Tratamiento de Información Multilingüe y Multimodal)*, 2011.

[4] M. Errecalde, D. Ingaramo, and P. Rosso. Itsa*: an effective iterative method for short-text clustering tasks. In *Proceedings of the 23th International Conference on Industrial, Engineering & Other Applications of Applied Intelligent Systems (IEA-AIE)*, 2010.

[5] D. Ingaramo, L. Cagnina, M. Errecalde, and P. Rosso. A particle swarm optimizer to cluster short-text corpora: a performance study. In *IBERAMIA: Workshop on Natural Language Processing and Web-based Technologies*, 2010.

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

[7] D. Ingaramo, M. Errecalde, and P. Rosso. A general bio-inspired method to improve the short-text clustering task. In *11th Intl. Conference on Intelligent Text Processing and Computational Linguistics*, 2010.

[8] Y. Lee, D. Strong, B. Kahn, and R. Wang. Aimq: A methodology for information quality assessment. *Information and Management*, 40(2), 2002.

[9] S. E. Madnick, R. Wang, Y. Lee, and H. Zhu. Overview and framework for data and information quality research. *ACM Journal of Data and Information Quality*, 1(1), 2009.

[10] H. Maurer, F. Kappe, and B. Zaka. Plagiarism - a survey. *Journal of Universal Computer Science*, 12(8), 2006.

[11] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2008.

[12] M. Potthast, A. Barrón, B. Stein, and P. Rosso. Cross-language plagiarism detection. *Language Resources and Evaluation, Special Issue on Plagiarism and Authorship Analysis*, 2010.

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

[14] F. Schlesinger, M. Errecalde, and G. Aguirre. An approach to integrate web services and argumentation into a bdi system. In *Proceedings of AAMAS*, pages 1371–1372, 2010.

[15] F. Schlesinger, E. Ferretti, M. Errecalde, and G. Aguirre. An Argumentation-based BDI Personal Assistant. In *23rd IEA-AIE*, LNAI. Springer, 2010.

[16] E. Stamatatos. Intrinsic plagiarism detection using character n-gram profiles. In *Proc. of the SEPLN 2009: Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse*, 2009.

[17] B. Stein. Principles of hash-based text retrieval. In *30th Annual International ACM SIGIR Conference*, 2007.

[18] B. Stein, S. Meyer zu Eissen, and M. Potthast. Strategies for retrieving plagiarized documents. In *30th International ACM SIGIR Conference*, 2007.

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

[20] S. Meyer zu Eissen and B. Stein. Intrinsic plagiarism detection. In *28th European Conference on IR Research*, 2006.