

Desarrollo de gestores de noticias incorporando técnicas basadas en la noción de confiabilidad

Cristian E. Briguez

Marcela Capobianco

Ana G. Maguitman

Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Departamento de Ciencias e Ingeniería de la Computación, Universidad Nacional del Sur,
Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)
Av. Alem 1253, (B8000CPB) Bahía Blanca, Argentina
e-mail: {ceb, mc, agm}@cs.uns.edu.ar

Resumen

La búsqueda y lectura de noticias en los distintos medios de información se ha convertido en una de las actividades más importantes dentro de la web. Como existen numerosas opciones de fuentes de noticias disponibles, es de suma importancia proveer mecanismos inteligentes para seleccionarlas y presentarlas de manera conveniente al usuario. Un factor importante que puede considerarse al momento de seleccionar noticias es la confiabilidad. El objetivo de esta investigación consiste en seguir avanzando en el diseño e implementación de gestores de noticias, mediante el desarrollo de nuevas técnicas inteligentes de procesamiento de noticias capaces de combinar nociones de relevancia, preferencia y confiabilidad. Dichas nociones serán abordadas utilizando tanto métodos cuantitativos como cualitativos, siendo estos últimos los que permitirán realizar un seguimiento sistemático y fundamentado de las noticias que interesan a los distintos usuarios. Para esto, se combinarán técnicas clásicas de administración de noticias con mecanismos de razonamiento lógicos tales como la argumentación rebatible.

Palabras clave: Gestión de Noticias, Argumentación, Búsqueda Contextualizada

Contexto

Esta línea de investigación se llevará a cabo dentro del ámbito del Laboratorio de Investigación y Desarrollo en Inteligencia Artificial (LIDIA), y está asociada a los siguientes proyectos de investigación:

- Diseño y Evaluación de Mecanismos de Búsqueda Contextualizada en Sistemas Centralizados y Distribuidos. PGI: 24/N029. Director: Ana G. Maguitman. Fuente de financiamiento: Universidad Nacional del Sur. 01/01/2011-31/12/2014.
- Soporte inteligente para el acceso a información Contextualizada en entornos centralizados y distribuidos. PIP: 11220090100863. Director: Ana G. Maguitman. Fuente de financiamiento: CONICET. 2010-2012.

Introducción

Buscar y leer noticias en diarios y otros medios de información que se encuentran en la Web se ha vuelto una de las actividades más importantes en este ámbito. Tanto la cantidad de usuarios como el tráfico en los principales buscadores de noticias (tales como Yahoo!

News, Google News) se ha incrementado de manera notoria en los últimos años. A esto debemos sumarle que también podemos encontrar noticias en las redes sociales (tales como Facebook, Twitter), las cuales han ganado en importancia y se han vuelto masivas. La gran cantidad de noticias online y la diversidad de fuentes de información disponibles refleja la necesidad que tienen los usuarios de estar informados y de obtener información y opiniones pluralistas.

La cantidad de fuentes de noticias disponibles en la Web es muy numerosa, por esto es de gran importancia proveer mecanismos inteligentes para seleccionarlas, combinarlas y presentarlas de manera conveniente al usuario. Los mayores esfuerzos de investigación en esta área se han concentrado en la *clasificación de noticias* de acuerdo a contenido de las noticias y el perfil de los usuarios. Sin embargo, debido a que las noticias publicadas en los periódicos no pueden ser verificadas en su totalidad, un factor importante que puede considerarse al momento de seleccionar noticias es la *confiabilidad*.

Existen aspectos importantes que deben ser tenidos en cuenta al momento de desarrollar un modelo realista de la noción de confiabilidad:

- La confiabilidad debe ser **justificada**.
- La confiabilidad es **rebatible**.
- La confiabilidad es **subjetiva**.

Los meta-buscadores de noticias generalmente pasan por alto la dinámica de la confiabilidad de noticias, o si es tenida en cuenta, la misma ha sido estudiada mediante técnicas puramente cuantitativas (*e. g.* [24]). Estas técnicas son limitadas en varios sentidos ya que dificultan la construcción de justificaciones para sostener la credibilidad de ciertas noticias. Dado que las técnicas cuantitativas no están dotadas de capacidad inferencial, mucha de la información

implícita no es revelada. Estas técnicas, por otra parte, son incapaces de tratar de manera formal con la naturaleza rebatible de la noción de confiabilidad.

Como ha ocurrido con otros aspectos de la Web, la gestión de noticias se ha vuelto una tarea altamente colaborativa. Sin embargo, la noción de confiabilidad es subjetiva y por lo tanto los modelos que tratan con dicha noción de manera objetiva no son totalmente realistas. En este caso, opiniones individuales de lectores confiables son más valiosas que las opiniones que provienen de agregar opiniones de un pool de lectores desconocidos.

En base a lo expuesto arriba, un paso importante para avanzar en el área del diseño e implementación de gestores de noticias es el desarrollo de nuevas técnicas inteligentes capaces de combinar nociones de relevancia, preferencia y confiabilidad. El trabajo propuesto consiste en la creación de un servicio de gestión de noticias que integre dichas técnicas y nos permita ver su comportamiento en distintas comunidades del mundo real.

Líneas de investigación y desarrollo

Esta línea de investigación se centrará sobre el área del diseño e implementación de gestores de noticias, mas concretamente en el desarrollo un sistema buscador y recomendador que integre nuevas técnicas de procesamiento de noticias creadas teniendo en cuenta la combinación de ciertas nociones, como relevancia, preferencia y confiabilidad.

En la actualidad, existe una gran cantidad de buscadores de noticias comerciales entre los que podemos destacar Google News [1], Yahoo! News [3], etc. Aunque ninguno de ellos da a conocer mediante publicaciones la manera en que ordenan las noticias, es evidente que toman en cuenta factores tales como la novedad, la fuen-

te a la que pertenece el artículo y la cantidad de veces que se repite en diferentes medios.

Una revisión de la literatura actual nos permite identificar una serie de desarrollos en el área de buscadores de noticias. NewsInEssence [27] es un sistema que busca y agrupa en clusters noticias relacionadas. QCS [12] es una herramienta que hace más eficiente la tarea de agrupar y categorizar los documentos en tópicos. En NewsJunkie [13] se propone y analiza un sistema que personaliza las noticias identificando las que son primicia en el contexto del usuario. Velthune [16] es un motor de búsqueda de noticias que extrae información tanto de la Web como de News feeds. Otro sistema, Compare&Contrast [21] es un sistema de razonamiento basado en casos que utiliza la Web como base de conocimiento para descubrir casos comparables de noticias.

Muchas técnicas han sido propuestas para organizar las noticias en tópicos. La mayoría tiene como objetivo reconocer, resumir y realizar un seguimiento de ciertos episodios (*e. g.* [25, 23, 20, 4]).

Otras propuestas han abordado los problemas de *credibilidad* y *sesgo* en los medios. En [24] se propone un método para evaluar la credibilidad de los artículos de noticias basado en la comparación de los contenidos de fuentes diversas. El servicio de noticias PolyNews [26] tiene como objetivo mitigar el efecto del sesgo introducido por diversos medios mediante la creación de múltiples puntos de vista. NewsTrust [2] es un servicio donde los usuarios pueden evaluar noticias, a sus redactores y a sus fuentes. Además, caben destacarse algunos modelos para el análisis de la propagación de la confiabilidad (*e. g.* [15, 17]) así como técnicas orientadas a analizar la diversidad y el sesgo de noticias de internet (*e. g.* [22]).

Otro área de especial interés se orienta a la detección de noticias de alto impacto (*e. g.* [18, 29]). Otro área de interés se enfoca en la representación de la reputación de los usuarios del

sistema como podemos ver en HARM [19].

En los últimos años la *argumentación rebatible* [11] evolucionó para convertirse en un mecanismo exitoso para modelar el razonamiento de sentido común cualitativo. En este sentido la *programación en lógica rebatible* (PLR) [14] y sus especializaciones la *programación en lógica rebatible posibilística* (PLRP) [5, 8] y la *programación en lógica rebatible con observaciones* (PLRO) [7]. Estos sistemas de razonamiento se han aplicado con éxito a diversas áreas, como *teorías de agencia* [6], *ingeniería de conocimiento* [9], *clustering* [10], y *negociación* [28]. En particular la PLRP fue diseñada para integrar el manejo de la incertidumbre en un sistema argumentativo. En trabajos posteriores [8] este sistema fue redefinido y optimizado para permitir incorporar percepciones y usar un mecanismo de optimización basado en el uso de conocimiento precompilado. Mediante este mecanismo el sistema podrá cumplir con los requerimientos de tiempo real necesarios para modelar el razonamiento en ambientes dinámicos.

La argumentación rebatible ha sido ya usada en ArgueNet [10] para diseñar un sistema de recomendaciones que clasifique los resultados de búsqueda de acuerdo a un conjunto de criterios de preferencia especificados por el usuario. Este acercamiento permite integrar en forma novedosa un motor de búsqueda tradicional con un framework argumentativo. El presente trabajo proyecta continuar esta línea de investigación, con especial énfasis en el tratamiento cualitativo de la noción de confiabilidad.

Resultados y Objetivos

Esta línea de investigación tiene por objetivo el desarrollo de un gestor de noticias que incorpore técnicas inteligentes capaces de combinar las nociones de relevancia, preferencia y confiabilidad. El trabajo a realizar consiste en

diseñar técnicas de procesamiento de noticias capaces de realizar un manejo cualitativo de dichas nociones. El objetivo específico es proponer un servicio que facilite el manejo cualitativo de la información. La meta es agregar un nivel adicional a los gestores de noticias básicos, donde se aplicarán nuevas heurísticas, se integrarán noticias basadas en fuentes antagónicas y donde será posible fundamentar posiciones contrapuestas. Esto nos permitirá integrar un sistema de recomendación de noticias con sistemas dialécticos, lo cual facilitará un seguimiento más sostenido y amplio de las noticias de interés para el usuario.

Formación de Recursos Humanos

Actualmente el equipo de trabajo de esta línea de investigación se encuentra compuesto por un becario de posgrado que accedió a una beca interna del conicet, su directora y su codirectora. Por otra parte se vincula con un grupo de trabajo sobre minería de datos compuesto por doctorandos e investigadores formados.

Referencias

- [1] <http://news.google.com/>.
- [2] <http://newstrust.net/>.
- [3] <http://news.yahoo.com/>.
- [4] James Allan, Rahul Gupta, and Vikas Khandelwal. Temporal summaries of news topics. In *SIGIR*, pages 10–18, 2001.
- [5] Teresa Alsinet, Carlos I. Chesñevar Lluís Godo, and Guillermo R. Simari. A logic programming framework for possibilistic argumentation: Formalization and logical properties. *Fuzzy Sets and Systems*, 159(10):208–228, 2008.
- [6] Daniel Bryant and Paul J. Krause. An implementation of a lightweight argumentation engine for agent applications. In *JELIA*, pages 469–472, 2006.
- [7] Marcela Capobianco, Carlos I. Chesñevar, and Guillermo R. Simari. Argumentation and the dynamics of warranted beliefs in changing environments. *Autonomous Agents and Multi-Agent Systems*, 11(2):127–151, 2005.
- [8] Marcela Capobianco and Guillermo R. Simari. A proposal for making argumentation computationally capable of handling large repositories of uncertain data. In *SUM*, pages 95–110, 2009.
- [9] Daniela V. Carbogim, David Robertson, and John Lee. Argument-based applications to knowledge engineering. *The Knowledge Engineering Review*, 15(2):119–149, 2000.
- [10] Carlos I. Chesñevar and Ana G. Maguitman. Arguenet: An argument-based recommender system for solving web search queries. In *Proc. of the 2nd IEEE Intl. IS-2004 Conference*, pages 282–287, 2004.
- [11] Carlos I. Chesñevar, Ana G. Maguitman, and Ronald P. Loui. Logical models of argument. *ACM Computing Surveys*, 32:337–383, 2000.
- [12] Daniel M. Dunlavy, John M. Conroy, and Dianne P. O’Leary. Qcs: A tool for querying, clustering, and summarizing documents. In *HLT-NAACL*, pages 11–12, 2003.
- [13] Evgeniy Gabrilovich, Susan T. Dumais, and Eric Horvitz. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *WWW*, pages 482–490, 2004.

- [14] Alejandro J. García and Guillermo R. Simari. Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming*, 4(1):95–138, 2004.
- [15] Ramanathan V. Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. pages 403–412, 2004.
- [16] Antonio Gulli. The anatomy of a news search engine. In *WWW (Special interest tracks and posters)*, pages 880–881, 2005.
- [17] Claudia Hess. Trust-based recommendations for publications: A multi-layer network approach. *TCDL Bulletin*, 2(2), 2006.
- [18] Yunliang Jiang, Cindy Xide Lin, and Qiaozhu Mei. Context comparison of bursty events in web search and online media. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*, pages 1077–1087, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.
- [19] Kalliopi Kravari and Nick Bassiliades. Harm: A hybrid rule-based agent reputation model based on temporal defeasible logic. In *RuleML*, pages 193–207, 2012.
- [20] Andrew J. Kurtz and Javed Mostafa. Topic detection and interest tracking in a dynamic online news source. In *JCDL*, pages 122–124, 2003.
- [21] Jiahui Liu, Earl Wagner, and Larry Birnbaum. Compare&contrast: using the web to discover comparable cases for news stories. In *WWW*, pages 541–550, 2007.
- [22] Qiang Ma and Masatoshi Yoshikawa. Topic and viewpoint extraction for diversity and bias analysis of news contents. In *Proceedings of the Joint International Conferences on Advances in Data and Web Management, APWeb/WAIM '09*, pages 150–161, Berlin, Heidelberg, 2009. Springer-Verlag.
- [23] Masaki Mori, Takao Miura, and Isamu Shioya. Topic detection and tracking for news web pages. In *Web Intelligence*, pages 338–342, 2006.
- [24] Ryosuke Nagura, Yohei Seki, Noriko Kando, and Masaki Aono. A method of rating the credibility of news documents on the web. In *SIGIR*, pages 683–684, 2006.
- [25] Ramesh Nallapati, Ao Feng, Fuchun Peng, and James Allan. Event threading within news topics. In *CIKM*, pages 446–453, 2004.
- [26] Souneil Park, Seungwoo Kang, and Junehwa Song. Polynews: Delivering multiple aspects of news to mitigate media bias. Technical Report KAIST, 2006.
- [27] Dragomir R. Radev, Jahna Otterbacher, Adam Winkel, and Sasha Blair-Goldensohn. Newsinsence: summarizing online news topics. *Commun. ACM*, 48(10):95–98, 2005.
- [28] Iyad Rahwan, Sarvapali D. Ramchurn, Nicholas R. Jennings, Peter McBurney, Simon Parsons, and Liz Sonenberg. Argumentation-based negotiation. *The Knowledge Engineering Review*, 18(3):343–375, 2003.
- [29] Aaron Sun, Daniel Dajun Zeng, and Hsinchun Chen. Burst detection from multiple data streams: a network-based approach. *Trans. Sys. Man Cyber Part C*, 40:258–267, May 2010.