Agentes inteligentes en ambientes dinámicos: minería de opiniones sobre en twitter

Pablo Kogan Fernando Buccella Lucas Mattar Sandra Roger

email: {pablo.kogan, fernando.buccella, roger}@fi.uncoma.edu.ar
lucas.mattar@gmail.com

Grupo de Investigación en Lenguajes e Inteligencia Artificial
Departmento de Teoría de la Computación
Facultad de Informática
Universidad Nacional del Comahue
Buenos Aires 1400 - (8300)Neuquén - Argentina

Resumen

El objetivo general de este Proyecto de Investigación es el estudio y desarrollo de técnicas de Inteligencia Artificial para dotar de inteligencia y conocimiento a agentes inmersos en mundos virtuales, interactivos y dinámicos, así como también el impacto que tienen las tecnologías del lenguaje humano (TLH) en la inclusión social. En estos escenarios, el razonamiento, la toma de decisiones, la planificación de acciones y el aprendizaje ocurren bajo restricciones de tiempo críticas y en intensa interacción con el usuario. Esta línea de investigación se centra en el desarrollo de una aplicación destinada al estudio y seguimiento de la opinión pública en Twitter sobre un tema determinado.

Palabras Clave: Agentes Inteligentes, Lenguaje Natural, Minería de opinión, análisis de sentimientos, accesibilidad.

Contexto

Este trabajo está parcialmente financiado por la Universidad Nacional del Comahue, en el contexto del proyecto de investigación *Agentes inteligentes en ambientes dinámicos*. El proyecto de investigación tiene prevista una duración de cuatro años, desde enero del 2013 hasta diciembre de 2016.

1. Introducción

El crecimiento de internet junto con el desarrollo de la Web 2.0 (Web Social) posibilita que personas de todo el mundo compartan y socialicen todo, y seguirá haciéndolo hasta puntos insospechados. Apenas recién nacida, la Web Social ya ha cambiado la forma en la que se genera y consume la información, y ha posibilitado el que todos seamos generadores de contenidos, que compartamos nuestro día a día con quienes queremos compartirlo, y sentemos las bases de lo que está por venir, la explotación inteligente de estos datos.

Junto con el desarrollo de la tecnología y el creciente acceso a la información, hemos sido testigos del nacimiento de un nuevo tipo de sociedad: la sociedad de la interactividad y comunicación [8].

La información textual disponible en la web podría ser categorizada en expresiones de hecho o de opinión. Las expresiones de hechos están relaciodas a entidades, eventos y sus propiedades. Por otro lado, las de opinión son

usualmente expresiones subjetivas que describen algún sentimiento o valoración sobre las personas, entidades, eventos y sus propiedades [6].

Uno de los *microblogging* más populares es Twitter. Los usuarios de Twitter pueden expresar las siguientes intenciones en sus tweets: charla diaria, conversaciones, intercambio de información, noticias y presentación de informes. Además, los usuarios de Twitter tienden a publicar las opiniones personales con respecto a ciertos temas y acontecimientos noticiosos. Una gran ventaja de estas opiniones es que se facilitan de manera libre y voluntariamente por los usuarios. Por lo tanto, los datos textuales de opiniones publicadas podría ser agregada y usada para medir la opinión pública implícita. Sin embargo, la elevada cantidad de opiniones generadas diariamente en aplicaciones de medios sociales, hace una evaluación humana de este contenido imposible de lograr. Por esta razón, estas opiniones textuales se suele evaluar mediante métodos computacionales

La minería de opinión o análisis de sentimientos se refiere a la aplicación de técnicas del campo del procesamiento del lenguaje natural, recuperación de información y clasificación de texto, para identificar y extraer información a partir de datos textuales subjetivos. Algunas de las tareas más importantes del campo son: distinguir entre la información objetiva y opiniones en fuentes de datos textuales, y para detectar los sentimientos en los textos dogmáticos al identificar si una tiene una opinión positiva o negativa relacionadas al tema tratado.

En este trabajo se pone principal énfasis en el seguimiento continuo de una temática en la web, junto con la determinación de los acontecimientos o hechos causales de variaciones en la opinión pública, siendo esto crucial a la hora de la toma de decisión. Brindando una información de gran valor estratégico que nos muestra una tendencia y/o comparativa de su valor mundial a través del tiempo.

2. Líneas de investigación y desarrollo

El proyecto de investigación *Agentes inteligentes en ambientes dinámicos* tiene varios objetivos generales. Por un lado, el de desarrollar conocimiento especializado en el área de Inteligencia Artificial. Además, se estudian técnicas de representación de conocimiento y razonamiento, junto con métodos de planificación y tecnologías del lenguaje natural aplicadas al desarrollo de sistemas multiagentes.

Específicamente, esta línea se centra en el estudio de un sistema multiagente en ambientes dinámicos para el seguimiento continuo de la opinión pública sobre un determinado tema de interés.

Las encuestas fueron tradicionalmente, la forma de obtener información acerca de la opinión pública, siendo estas estáticas en un tiempo discreto. A diferencia de este tipo de encuestas, este trabajo está enfocado en realizar un seguimiento continuo de la opinión pública. Está opinión está expresada públicamente en diferentes sitios de la web. Teniendo este corpus a disposición el proceso continua realizando una clasificación de la información obtenida acerca de la temática a analizar. Por ejemplo si la temática a analizar es el “asignación universal por hijo” se pueden buscar los comentarios de las noticias relacionadas con este tema, y clasificarlos si están a favor o en contra.

El objetivo de esta investigación es desarrollar una herramienta para hacer este proceso de forma automática.

La arquitectura básica de nuestro sistema multiagente [7] está dividida en cuatro agentes principales: Agente Buscador, Agente Filtrador, Agente Analizador y Agente Compositor.

El primer agente que entra en juego es el **Agente Buscador**, el cual tiene tres tareas principales: análisis de la entrada o consulta, búsqueda sobre la web, y finalmente almacenar lo buscado en una base de datos.

Se producen además dos procesos. En el primero, se realiza una desambigüación de la entrada en el caso de ser necesario. Por ejemplo si se está buscando a Riquelme, se pro-

ducirá una desambigüación entre el jugador de fútbol argentino "Juan Román Riquelme"; la modelo paraguaya "Larissa Riquelme"; el niño holandés cuyo nombre es "Riquelme Van Gool", en homenaje al jugador de fútbol; entre otros. El segundo proceso, trata también con la consulta pero en el sentido de producir la relajación de la entrada, esto quiere decir que si estamos buscando a "Cristina Fernandez de Kirchner" también se considere, por ejemplo, al término "presidenta de la Argentina" como equivalente. En este sentido debemos identificar un algoritmo óptimo para la construcción de la entrada y sus limitaciones. Usando las consultas correctas se podrá encontrar las sentencias adecuadas en el proceso de recuperación.

El proceso de búsqueda es realizado en diferentes ámbitos:

- *Búsqueda en microblogs conocidos:* Los *microblogs* se han convertido en una herramienta muy popular entre los usuarios de internet. Millones de mensajes aparecen diariamente en los sitios más populares de *microblogging* como *Twitter, Facebook, Tumblr*, etc. Los autores de estos mensajes escriben acerca de su vida, comparten sus opiniones sobre una variedad de temas y discuten sobre estos. Como el formato de los mensajes es libre y de fácil acceso, los usuarios tienden a modificar su forma de comunicación de blogs y listas de correos tradicionales a servicios de *microblogging*.

- *Comentarios de noticias de diarios (La Nación, Clarín, etc.):* La proliferación de los diarios en su versión *on-line*, posibilitan a los lectores la opción de comentar las noticias, con el objetivo de hacer al diario más interactivo. Esta fuente de información es muy rica en contenido y en opinión. La búsqueda sobre los comentarios no es tan trivial como la anterior. Esta se realiza a través de un robot web que va navegando las noticias relacionadas con el tema y almacenando los comentarios.

- *Búsqueda en la web a través de buscadores:* aprovechando el resultado que arrojan los buscadores se realiza un robot web que navega los *links* y devuelve resultados de blogs, listas de correos públicos y noticias de sitios poco conocidos.

Toda la información obtenida se almacena en una base de datos con toda la información que se puede obtener de la persona que publica su opinión.

El **Agente Filtrador** se encarga de descartar todos los datos del corpus que no sirven, como por ejemplo entradas duplicadas, entradas que no demuestran sentimientos, etc..

El **Agente Analizador** se encarga de clasificar las entradas del corpus en sentimientos. Inicialmente comenzaremos a trabajar con una ontología de dos sentimientos: "amor" y "odio". Este agente es el encargado de realizar un proceso de entrenamiento sobre análisis de sentimientos. Esta tarea es realizada con la herramienta Weka[1] e inicialmente utilizado el clasificador *Support Vector Machine* (SVM) dado su relativo éxito en el tratamiento del lenguaje natural. Posteriormente se realizará un estudio comparativo más profundo sobre otros clasificadores.

Finalmente, el **Agente Compositor** es el encargado de componer los resultados obtenidos por el agente analizador en un lapso de tiempo determinado. El factor tiempo en conjunto con los resultados son los puntos mas importantes a analizar. Los resultados obtenidos podrían modificar el comportamiento del agente buscador antes de comenzar un nuevo ciclo.

2.1. Baseline inicial

En la primera versión del sistema, se creó un *baseline* con el cual comparar posteriormente diferentes mejoras en las diferentes etapas del sistema. Dicho *baseline* se definió de la siguiente manera.

El *agente buscador* toma Twitter y se encarga de hacer una recuperación o filtrado de tweets sobre un tema específico, almacenando

[1] www.cs.waikato.ac.nz/ml/weka/

los mismos en una base de datos con información del usuario emisor del tweet. No se produce, además, ninguna consideración relacionada al tweet. Es decir, no se tiene en cuenta si el tweet es de opinión o informativo. No se produce ningún proceso de desambigüación ni de relajación de la entrada. En esta instancia el agente buscador sólo conoce la entidad sobre la cual se desea realizar un análisis de opinión y el rango de fechas que se desea utilizar para medir la opinión de los usuarios acerca del tema planteado.

El *Agente Filtrador* solamente se encarga de descartar los datos redundantes del corpus, como por ejemplo entradas duplicadas. En este sentido, se ha decidido realizar esta eliminación dado que no se utiliza ninguna información relacionada al usuario emisor de la opinión.

El **Agente Analizador** encargado de clasificar las entradas del corpus en sentimientos, comenzará a trabajar con una ontología de dos sentimientos: "amor" / "odio"o "positivo"El recurso lingüístico utilizado aquí es el diccionarios / "negativo". Este agente es el encargado de realizar un proceso de entrenamiento sobre análisis de sentimientos. En este caso el algoritmo será considerado positivo si tiene un número mayor de palabras positivas que negativas.

Finalmente, el **Agente Compositor** compondrá los resultados obtenidos por el agente analizador en el lapso de tiempo introducido como entrada en el agente buscador.

2.2. Enfoque basado en diccionarios

Hay dos enfoques relacionados para encontrar palabras de opinión: el enfoque basados en corpus [3, 9, 1, 10] y el enfoque basados en diccionarios [4, 5, 2]. En este punto se pretende realizar un enfoque basado en diccionarios mediante la creación de diccionarios de sentimientos para ser utilizados por el sistema.

3. Resultados esperados

El objetivo de este sistema es lograr una herramienta web accesible. De esta manera, el usuario puede proponer una temática para analizar el comportamiento de la opinión pública sobre dicho tema. Actualmente, se está trabajando en producir resultados que sirvan de base de comparación a futuros análisis y mejoras. En este sentido, se pretende analizar diferentes fuentes de búsqueda, algoritmos de clasificación, herramientas lingüísticas, etc. para un mejor desempeño del sistema.

4. Formación de Recursos Humanos

Este proyecto cuenta con dos integrantes del proyecto, un becario alumno de la Universidad Nacional del Comahue y un tesista de grado. Se espera lograr, a lo largo de este año, la incorporación de un becario y un tesista de carrera de grado.

Finalmente, es constante la búsqueda hacia la consolidación como investigadores de los miembros más recientes del grupo

Referencias

[1] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the Conference on Web Search and Web Data Mining*. (WSDM-2008), 2008.

[2] A. Esuli and F. Sebastiani. Determining the orientation of terms through gloss classification. In *Proceedings of ACM International Conference on Information and Knowledge Management*. (CIKM-2005), 2005.

[3] V. Hatxivassiloglou and K. McKeown. Predicting the semantic orientation of adjetives. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*. (ACL-1997), 1997.

[4] M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of SIGKDD International Conference and Knowledge Discovery and Data Mining*. 2008.

[5] S. Kim and E. Hovy. Determining the sentiment of opinions. In *Proceedings of International Conference on Computational Linguistics*. (COLING-2004), 2004.

[6] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[7] S. Roger and P. Kogan. Agentes inteligentes en ambientes dinámicos: minería de opiniones sobre en twitter. In *XV Workshop de Investigadores en Ciencias de la Computación*. (WICC-2013), 2013.

[8] M. Wiberg. *The Interaction Society: Theories, Practice and Supportive Technologies*. Information Science Publishing, 2004.

[9] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizingcontextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing*. (HLT/EMNLP-2005), 2005.

[10] Y. Wu and M. Wen. Disambiguating dynamic sentiment ambiguous adjetives. In *Proceedings of 23rd International Conference on Computational Linguistics*. (Coling 2010), 2010.