Análisis de opinión como un sistema multiagente distribuido

Pablo Kogan Sandra Roger

email: {pkogan,sroger}@uncoma.edu.ar

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 fundamental de este proyecto es el desarrollo de conocimiento especializado en el área de Inteligencia Artificial Distribuida, estudiando técnicas de representación del conocimiento y razonamiento, junto con métodos de planificación y tecnologías del lenguaje natural aplicadas al desarrollo de sistemas multiagentes.

Esta línea de investigación se centra en el desarrollo de una aplicación destinada al estudio y seguimiento de la opinión pública sobre un tema determinado.

El crecimiento de internet junto con el desarrollo de la Web 2.0 (Web Social) posibilita que personas de todo el mundo compartan información global. Millones de mensajes aparecen diariamente en los sitios más populares de *microblogging*, comentarios de noticias en diarios web, blogs, etc. Los autores de estos mensajes escriben acerca de sus vidas, comparten sus opiniones sobre una variedad de temas y discuten sobre estos. Toda esta información que los usuarios generan en las publicaciones acerca de los productos que utilizan o visión política y religiosa, se vuelve un recurso de gran valor para el análisis de opiniones y sentimientos de la opinión pública.

El estudio de estos temas mediante un seguimiento continuo junto con la determinación sobre los acontecimientos o hechos causales de variaciones en la opinión pública son cruciales a la hora de tomar una decisión. Tanto a nivel de consumidor como de proveedor esta información tiene un gran valor estratégico, que les brinda una tendencia y/o comparativa del valor mundial a través del tiempo.

Palabras Clave: Agentes inteligentes, Sistemas multiagentes, Procesamiento del lenguaje natural, Opinion Mining, Análisis de sentimientos.

Contexto

Este trabajo está parcialmente financiado por la Universidad Nacional del Comahue, en el contexto del proyecto de investigación *Sistemas Multiagentes en Ambientes Dinámicos: Planificación, Razonamiento y Tecnologías del Lenguaje Natural*. El proyecto de investigación tiene prevista una duración de tres años, ha comenzado en enero del 2010 y finaliza en diciembre de 2012.

1. Introducción

La información textual disponible en la web podría ser categorizada en expresiones de hecho o de opinión. Las expresiones de hechos están relaciondas a entidades, eventos y sus propiendades. 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].

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 [9]. En este nuevo contexto, el papel de los sentimientos expresados en la web se han vuelto crucial. Las personas expresan las emociones que determinan ciertos hechos. Por otra parte, otras personas que tienen acceso a estas expresiones, las transforman bajo sus propias influencias en otras expresiones.

El crecimiento de lugares on-line en donde las personas pueden expresar sus opiniones abre un nuevo campo de investigación: la minería de opinión (*Opinion Mining* -OM-).

La investigación en OM es una disciplina reciente concerniente a la recuperación de opiniones expresadas en un documento y no sobre el tema del mismo como es el caso de la recuperación de información. Más específicamente está relacionado con la opinión de un autor expresada en un documento o texto, como ser blogs, micro-blogs, noticias, comentarios, etc. [3]. Este análisis de sentimientos conlleva algunos desafíos, entre ellos, determinar si cada segmento de texto (sentencia, párrafo o sección) es una opinión o no; identificar quién expresa la opinión (una persona, organización, etc.) y determinar si la opinión es positiva, negativa o neutra, o de acuerdo a una cierta taxonomía preestablecida.

Teniendo en cuenta la riqueza del lenguaje humano y su gran poder expresivo y ambigüedad inherente al mismo, el problema de la clasificación de sentimientos no es trivial.

Las aplicaciones en donde puede ser crucial el análisis de los sentimientos pueden ser:

- Resumenes de opiniones [2].
- Opiniones de libros o películas [4, 5]
- Análisis de opiniones políticas: Análisis de candidatos políticos [1, 7], e-goverment [8] para analizar impacto de desiciones.
- Análisis del impacto de productos y marcas [1].

Los recursos lingüísticos para OM definen algunas propiedades relacionadas a los sentimientos. Los avances sobre este tópico tratan con tres tareas principales:

- *Determinación de la orientación del término*: positivo, negativo, neutro.
- *Determinación de la subjetividad de un término*, si un término tiene una naturaleza subjetiva u objetiva: verde, alto, líquido, etc.
- *Determinación de la fuerza de la determinación del término (orientación o subjetividad)*, como el grado de positividad o negatividad del término.

Las investigaciones sobre OM han tomado tres líneas de investigación interrelacionadas [6]

- Desarrollo de recursos lingüísticos para el análisis de sentimientos tal como corpus léxico anotado manualmente;
- Implementación de diferentes algoritmos para el análisis del texto y clasificación de acuerdo a su orientación semántica y subjetiva;
- Extracción de opiniones del texto, incluyendo diferentes tipos de relaciones con contenido asociado.

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 *Sistemas Multiagentes en Ambientes Dinámicos: Planificación, Razonamiento y Tecnologías del Lenguaje Natural* tiene varios objetivos generales. Por un lado, el de desarrollar conocimiento especializado en el área de Inteligencia Artificial Distribuida. 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.

Específicamenete, 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 encontra.

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

La Figura 1 muestra la arquiterctura básica del sistema multiagentes destinada al análisis de opinión. La misma 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.

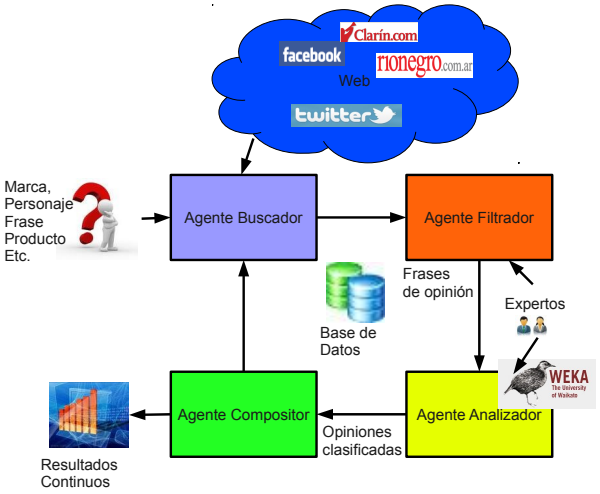

Figura 1: Arquitectura de agentes para análisis de opinión

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 producirá 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*. Dado el gran crecimiento de las publicaciones por parte de los usuarios acerca de los productos que utilizan o visión política y religiosa, los sitios de *microbloggin* se vuelven un recurso de valor para las opiniones y sentimientos de la opinión pública. Estos sitios brindan herramientas de búsqueda a través de un *web service* por lo cual hace que esta tarea sea bastante simple.

- *Comentarios de noticias de diarios (La Nación, Clarín, etc.):* La proliferación de los diaros 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.

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.

Referencias

[1] C. G. Akcora, M. A. Bayir, M. Demirbas, and H. Ferhatosmanoglu. Identifying breakpoints in public opinion. *1st Workshop on Social Media Analytics (SOMA 10)*, 2010.

[2] A. Bossard, M. Généreux, and T. Poibeau. Cbseas, a summarization system integration of opinion mining techniques to summarize blogs. In *Proceedings of the 12th Conference of the European Chapter of the*

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

Association for Computational Linguistics: Demonstrations Session, EACL '09, pages 5–8, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[3] B. Liu. Sentiment analysis and subjectivity. In N. Indurkhya and F. J. Damerau, editors, *Handbook of Natural Language Processing, Second Edition*. CRC Press, Taylor and Francis Group, Boca Raton, FL, 2010. ISBN 978-1420085921.

[4] S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. Mining product reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 341–349, New York, NY, USA, 2002. ACM.

[5] B. Pang and L. Lee. Thumbs up? sentiment classification using machine learning techniques. In *In Proceedings of EMNLP*, pages 79–86, 2002.

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

[7] M. J. Silva, P. Carvalho, L. Sarmento, E. Oliveira, and P. Magalhães. The design of OPTIMISM, an opinion mining system for portuguese politics. In *New Trends in Artificial Intelligence: Proceedings of EPIA 2009 - Fourteenth Portuguese Conference on Artificial Intelligence*. Universidade de Aveiro, Oct. 2009.

[8] G. Stylios, D. Christodoulakis, J. Besharat, M.-A. Vonitsanou, I. Kotrotsos, A. Koumpouri, and S. Stamou. Public opinion mining for governmental decisions. In F. Bannister, editor, *ECEG Conference Issue*, volume 8, pages 202 – 213, 2010.

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