Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 139-142 recibido 10-04-14 revisado 02-06-14 aceptado 10-06-14

Tratamiento inteligente de la información para ayuda a la toma de decisiones

Intelligent information processing to support decision-making

Sonia Vázquez, Elena Lloret, Fernando Peregrino, Yoan Gutiérrez, Javier Fernández, José Manuel Gómez
Universidad de Alicante
Carretera San Vicente del Raspeig s/n 03690, Alicante, España
{svazquez, elloret, fsperegrino, ygutierrez, javifm, jmgomez}@dlsi.ua.es

Resumen: Proyecto emergente centrado en el tratamiento inteligente de información procedente de diversas fuentes tales como micro-blogs, blogs, foros, portales especializados, etc. La finalidad es generar conocimiento a partir de la información semántica recuperada. Como resultado se podrán determinar las necesidades de los usuarios o mejorar la reputación de diferentes organizaciones. En este artículo se describen los problemas abordados, la hipótesis de trabajo, las tareas a realizar y los objetivos parciales alcanzados.
Palabras clave: Minería web, tratamiento de información textual, PLN

Abstract: This project is focused on intelligent information processing using different sources such as micro-blogs, blogs, forums, specialized websites, etc. The goal is to obtain new knowledge using semantic information. As a result we can determine user requirements or improve organizations reputation. This paper describes the problems faced, working hypothesis, tasks proposed and goals currently achieved.
Keywords: Web mining, textual information processing, NLP

1 Datos del proyecto

Proyecto dirigido por Sonia Vázquez, miembro del Grupo de Procesamiento del Lenguaje y Sistemas de Información (GPLSI) de la Universidad de Alicante. Financiado por la Universidad de Alicante (GRE12-44) dentro del programa de ayudas a proyectos emergentes. Inicio 01/09/2013 (duración 2 años).

Contacto

Email: svazquez@dlsi.ua.es
Teléfono: 965903400 ext. 2947
Dpto. de Lenguajes y Sistemas Informáticos
Universidad de Alicante
Carretera San Vicente del Raspeig s/n,
03690, Alicante, España.

2 Introducción

Actualmente, las opiniones de los consumidores sobre diferentes tipos de productos y servicios están disponibles en Internet en diferentes lugares pudiendo ser expresadas a través de redes sociales, foros, portales especializados, blogs personales, etc. Mediante este tipo de información los usuarios que deseen adquirir un nuevo producto o servicio pueden consultar las críticas realizadas por diferentes usuarios acerca de ciertas características concretas. En muchas ocasiones, las críticas no son realizadas por expertos sino por usuarios reales que han probado el producto o el servicio y dan su opinión desde su punto de vista particular. En estos casos, las críticas pueden ser incompletas y centrarse únicamente en ciertos aspectos, de forma que el posible comprador o nuevo usuario recibe la información sesgada y debe buscar en diferentes lugares hasta conseguir una visión más completa de las características del producto. En ocasiones, la información que ofrecen algunos portales de Internet donde los usuarios pueden opinar acerca de diferentes productos viene acompañada de puntuaciones que indican el grado de utilidad de esa crítica, pudiendo seleccionar aquellas que tengan mayor puntuación para facilitar la toma de decisiones (Amazon, Ciao).

En los últimos años, se han realizado diversos estudios enfocados a la detección de la subjetividad en los textos llegando incluso a

ISSN 1135-5948

determinar la intensidad de dichas opiniones (disgusto, agrado, ironía, felicidad, etc). La información relativa a opiniones o críticas por parte de diferentes usuarios aparece dispersa en Internet por lo que es necesario establecer mecanismos que permitan una correcta búsqueda, recopilación y utilización de la misma. La tarea de descubrir conocimiento útil a través de información procedente de Internet es conocida como Web mining.

Dado el inminente interés generado por tratar la información subjetiva presente en la web se han desarrollado diversos recursos que permiten detectar el grado de afectividad presente en los textos. Estos recursos sirven como base de conocimiento para diferentes tipos de sistemas.

Manejar toda la información disponible y conseguir obtener una visión general de las opiniones de los usuarios es una tarea muy compleja y que requiere de diversas técnicas de PLN: detección de la informalidad, geolocalización, generación de resúmenes automáticos, resolución de la ambigüedad, etc.

3 Objetivos del proyecto

El objetivo principal de este proyecto es el tratamiento inteligente de la información textual procedente de diversas fuentes tales como micro-blogs, blogs, foros, portales especializados, etc. Mediante el uso de técnicas de PLN se extraerán de forma automática las principales características sobre un producto o un servicio y se recuperará información (opiniones de usuarios) relativa a estas características. Entre los problemas que se deben resolver se encuentran la detección de ironía o el sarcasmo, la ambigüedad, la geolocalización, la informalidad en los textos y la generación automática de resúmenes.

La extracción de características se realizará atendiendo a diferentes dominios de aplicación como por ejemplo el sector tecnológico, turístico, económico, etc. En cuanto al tema de la geolocalización se determinará la distribución geográfica de las páginas que hablan sobre un producto o un servicio para determinar en qué zonas ha tenido mejor acogida y en cuáles no. En este caso, se podrán restringir las búsquedas a ciertas zonas, como por ejemplo, la provincia de Alicante. La informalidad en los textos será tratada para transformar elementos de carácter informal en sus correspondientes implicaciones formales de manera que no se pierda el contenido semántico implícito en la descripción inicial. Además, las tareas de detección de opiniones y generación de resúmenes automáticos completarán el sistema para proporcionar unos resultados que generen nuevo conocimiento.

El interés de este proyecto viene determinado por la necesidad de disponer de información veraz de forma inmediata proveniente de diferentes fuentes y que abarque un amplio abanico de posibilidades. De esta forma tanto usuarios como entidades u organizaciones, tendrán el conocimiento necesario para poder tomar decisiones lo más acertadas posibles. De forma que un usuario pueda decantarse por la compra de un producto o servicio o una organización centre sus esfuerzos en mejorar ciertas partes de su imagen, sus productos o servicios.

4 Hipótesis de trabajo

Esta investigación se centra en la hipótesis de que la información procedente de las opiniones y críticas presentes en Internet puede ser utilizada para mejorar la reputación de organizaciones o determinar las necesidades de los usuarios para la creación de nuevos productos.

Debido al rápido crecimiento de la información en Internet los usuarios potenciales de un producto determinado tienen muchas dificultades a la hora de realizar una elección. Cualquier modificación en el producto, en las políticas de la empresa o en la atención al usuario puede variar de forma sustancial la opinión de los usuarios y posibles compradores. Poseer la información de las reacciones del público en general puede aportar beneficios a las diferentes partes implicadas tanto para la toma de decisiones por parte de los usuarios como para la mejora de los productos y servicios por parte de las organizaciones. Por ello, es necesario establecer qué necesidades se han visto cubiertas y cuáles quedan pendientes para poder desarrollar nuevos productos que satisfagan a la mayoría de usuarios. Mediante la utilización de técnicas de PLN se conseguirá generar conocimiento a partir de la información semántica, la localización, las opiniones, etc. De esta forma, se facilitará la toma de decisiones por parte de usuarios y organizaciones.

5 Tareas a desarrollar

Para la consecución del proyecto será necesario completar el conjunto de tareas y subtareas que se mencionan a continuación:

Análisis del problema

Se analizarán las distintas aproximaciones existentes para la extracción automática de características, detección del foco geográfico, informalidad en los textos, detección de opiniones y generación automática de resúmenes y conocimiento. Sobre esta base teórica se investigarán nuevas técnicas para la mejora de cada una de las actividades implicadas en el sistema.

Estudio de técnicas para la extracción automática de características

Se determinarán las características más relevantes relacionadas con un producto o servicio. Se estudiarán las aproximaciones actuales de extracción automática de características para desarrollar un sistema de dominio abierto que pueda ser aplicado a diferentes sectores (Dave, Lawrence, y Pennock, 2003), (Gamon et al., 2005).

Estudio de técnicas para la recuperación y extracción de información

Se realizará un proceso de recuperación y extracción de información cuya finalidad sea la de seleccionar de forma automática los textos relativos a esas características (Baeza-Yates y Ribeiro-Neto, 2011), (Manning, Raghavan, y Schütze, 2008). En este caso, se utilizarán diversas fuentes de información tales como micro-blogs, blogs, foros y portales especializados. Se estudiarán las técnicas actuales para recuperación y extracción de información tratando de adaptar dichas técnicas a nuestro ámbito de aplicación, resolviendo diferentes tipos de problemas como la detección de entidades entre otros (Kozareva, Vázquez, y Montoyo, 2007).

Estudio de técnicas para la detección de opiniones

Se realizará una clasificación basada en la opinión de los usuarios acerca de las diferentes características o servicios de los productos. Se realizará un estudio de las técnicas actuales adaptando las mejores aproximaciones a cada entorno específico. Técnicas basadas en la combinación de características semánticas y adjetivos (Hatzivassiloglou y Wiebe, 2000), métodos basados en conocimiento (Gutiérrez, Vázquez, y Montoyo, 2011), técnicas basadas en la extracción de términos relevantes que definan la polaridad (Zubaryeva y Savoy, 2010), métodos basados la proximidad semántica entre conceptos (Balahur y Montoyo, 2009).

Estudio de técnicas para la detección de la informalidad

Se tratará la informalidad de los textos sobre opiniones y críticas de una forma específica (Mosquera y Moreda, 2012). Se determinarán las aproximaciones más adecuadas para tratar textos provenientes de diversas fuentes como: micro-blogs, blogs, foros y portales especializados (Buscaldi y Rosso, 2008).

Estudio de técnicas para la geolocalización

Se realizará un estudio sobre diferentes técnicas de desambiguación (Vázquez, Montoyo, y Kozareva, 2007) y detección de entidades. Debido a que un mismo topónimo puede pertenecer a diferentes lugares es crucial establecer de qué lugar concreto se está hablando (Peregrino, Tomás, y Llopis, 2011).

Estudio de técnicas para la generación de resúmenes y nuevo conocimiento

El objetivo de esta tarea es la determinar qué información será la más relevante para un usuario y sintetizarla en forma de un pequeño texto. Debido a la gran cantidad de información existente los usuarios son incapaces de manejar de manera eficiente dicha información. Por tanto, a partir de toda la información recopilada y clasificada según el grado de aceptación de los usuarios se estudiarán nuevas técnicas para la generación de conocimiento que permitan sintetizar de forma coherente y veraz la gran cantidad de información relacionada con un producto o servicio concreto (Barzilay y McKeown, 2005), (Lloret y Palomar, 2013).

6 Situación actual del proyecto

Dentro de las tareas antes mencionadas, hasta el momento se ha desarrollado un recurso que relaciona diferentes bases de datos léxicas y ontologías junto con el grado de afectividad (Gutiérrez et al., 2011): SentiWordnet, WordNet, SUMO, WordNet Affect. Además, se ha mejorado un sistema de desambiguación basado en conocimiento para adecuarlo a las necesidades del proyecto. Y por último, se han estudiado diversas técnicas para mejorar la generación automática de resúmenes.

Bibliografía

Baeza-Yates, Ricardo A. y Berthier A. Ribeiro-Neto. 2011. *Modern Information Retrieval - the concepts and technology behind search, Second edition*. Pearson Education Ltd., Harlow, England.

Balahur, Alexandra y Andrés Montoyo. 2009. A semantic relatedness approach to classifying opinion from web reviews. *Procesamiento del Lenguaje Natural*, 42.

Barzilay, Regina y Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Comput. Linguist.*, 31(3):297–328, Septiembre.

Buscaldi, Davide y Paolo Rosso. 2008. Geo-wordnet: Automatic georeferencing of wordnet. En *LREC*. European Language Resources Association.

Dave, Kushal, Steve Lawrence, y David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. En *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, páginas 519–528, New York, NY, USA. ACM.

Gamon, Michael, Anthony Aue, Simon Corston-Oliver, y Eric Ringger. 2005. Pulse: Mining customer opinions from free text. En *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis*, IDA'05, páginas 121–132, Berlin, Heidelberg. Springer-Verlag.

Gutiérrez, Yoan, Antonio Fernández Orquín, Sonia Vázquez, y Andrés Montoyo. 2011. Enriching the integration of semantic resources based on wordnet. *Procesamiento del Lenguaje Natural*, 47:249–257.

Gutiérrez, Yoan, Sonia Vázquez, y Andrés Montoyo. 2011. Sentiment classification using semantic features extracted from wordnet-based resources. En *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11, páginas 139–145, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hatzivassiloglou, Vasileios y Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. En *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING '00, páginas 299–305, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kozareva, Zornitsa, Sonia Vázquez, y Andrés Montoyo. 2007. Multilingual name disambiguation with semantic information. En *TSD*, páginas 23–30.

Lloret, Elena y Manuel Palomar. 2013. Compendium: a text summarisation tool for generating summaries of multiple purposes, domains, and genres. *Natural Language Engineering*, 19(2):147–186.

Manning, Christopher D., Prabhakar Raghavan, y Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Mosquera, Alejandro y Paloma Moreda. 2012. The study of informality as a framework for evaluating the normalisation of web 2.0 texts. En *Proceedings of the 17th International Conference on Applications of Natural Language Processing and Information Systems*, NLDB'12, páginas 241–246, Berlin, Heidelberg. Springer-Verlag.

Peregrino, Fernando S., David Tomás, y Fernando Llopis. 2011. Map-based filters for fuzzy entities in geographical information retrieval. En *NLDB*, páginas 270–273.

Vázquez, Sonia, Andrés Montoyo, y Zornitsa Kozareva. 2007. Word sense disambiguation using extended relevant domains resource. En *IC-AI*, páginas 823–828.

Zubaryeva, Olena y Jacques Savoy. 2010. Opinion detection by combining machine learning & linguistic tools.