ARTÍCULO ACEPTADO

Sistemas Web colaborativos para la recopilación de datos bajo el paradigma de ciencia ciudadana

por **Alejandro Molina Villegas**

En algunas áreas científicas, la obtención de datos requiere un enorme trabajo manual y pocas veces se cuenta con los recursos necesarios para cubrir esta necesidad. Sin embargo, la Web nos ofrecen una alternativa para superar esta situación. Muchos científicos han decidido involucrar voluntarios no-expertos en actividades ligadas con la recolección de datos para la investigación. Algunos de ellos han recabado enormes cantidades de datos a través de sistemas Web colaborativos. *Ciencia ciudadana* (en Inglés conocida como Citizen science, crowd science o networked science) es el término adoptado para referirse a este tipo de actividades que, entre otras ventajas, permite obtener datos de manera rápida y a bajo costo. En este artículo, se presentan algunos proyectos bajo este modelo y se describen dos sistemas que hicieron posible la anotación manual de una gran cantidad de textos, para las áreas de resumen automático de documentos y análisis de opinión en microblogs.

El paradigma de ciencia ciudadana es una alternativa para incorporar gente en proyectos científicos que requieren la recolección de grandes cantidades de datos

El paradigma de ciencia ciudadana

La ciencia ciudadana consiste en involucrar voluntarios no-expertos en tareas científicas que no requieren ninguna pericia de la materia en cuestión. Las razones para implicar ciudadanos en actividades científicas son varias y no se contraponen unas con otras: economizar recursos humanos, aumentar la cantidad y la rapidez de procesamiento de los datos, o simplemente, acercar la gente a la ciencia. Uno de los objetivos del proyecto Neighborhood Nestwatch (Evans 2005), fue concientizar a los habitantes de Washington acerca de la biodiversidad de su entorno. El proyecto Galaxy Zoo, descrito en (Fortson 2011), invita a los usuarios a clasificar galaxias usando las imágenes capturadas por un satélite. Esto, que parecía al inicio una meta difícil de conseguir, ha permitido concentrar una base de datos con 900 000 galaxias clasificadas manualmente por voluntarios de todo el planeta. Otro proyecto similar, permitirá la transcripción de millones de papiros en griego antiguo descubiertos por arqueólogos británicos. En (Nurmikko 2012) consideran que permitiendo al público transcribir una letra a la vez acelerarán un proceso que podría haber durado mucho años.

El concepto de ciencia ciudadana no es nuevo pero ha cobrado auge en el último decenio, probablemente debido a la proliferación de sitios colaborativos en la Web. Muchos investigadores se han inclinado por este modelo. En las áreas involucradas con tecnologías del lenguaje representa una alternativa para la anotación de documentos a gran escala. Esta práctica es conocida como *anotación de corpus* y consiste en marcar, de manera masiva, algunas partes de una colección de textos (un corpus) con indicaciones especiales según el interés de la investigación. Por ejemplo, se pueden indicar en una colección de libros todas las apariciones de verbos en infinitivo.

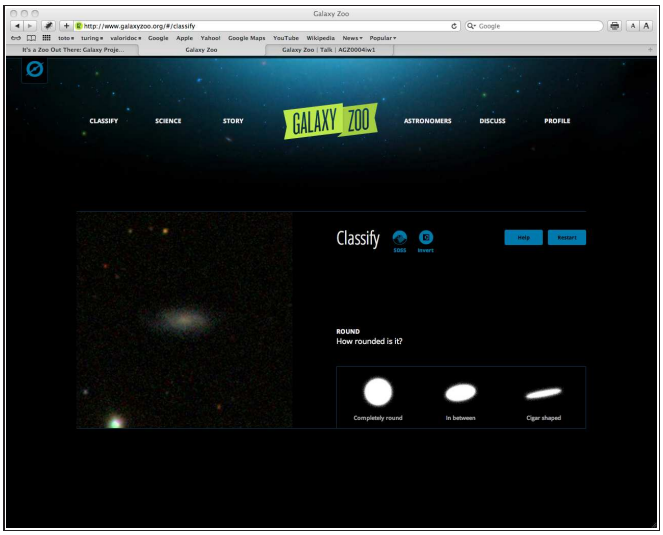

Figura 1. Clasificación de galaxias para proyecto Galaxy Zoo

El proyecto LEARNER (Chklovski 2003), permite al cybernauta común convertirse en un maestro que transmitirá a una computadora su conocimiento mediante sesiones de preguntas y respuestas. En otro proyecto, descrito en (Chamberlain 2009), se desarrolló un juego de video en el que el usuario (un detective) debe descubrir un misterio desambigüando textos con referen-

 ISSN 2007-0691

cias anafóricas. En el resto del artículo presentamos dos proyectos para los cuales se diseñaron campañas de anotación para la recopilación de datos. El primer proyecto trata acerca de la generación automática de resúmenes de documentos mientras que el segundo trata acerca del análisis de la opinión en microblogs.

Recopilación de datos para un proyecto de resumen automático

Hoy en día, existen programas capaces de identificar, con gran precisión, cuáles son las oraciones más importantes de un texto. A este método se le conoce como *resumen extractivo*. Sin embargo, existe el inconveniente de que una gran parte de la información secundaria aparece al interior de dichas oraciones. Una técnica recientemente estudiada consiste en *comprimir* las oraciones en el texto. Esto es, eliminar palabras de una oración con la intención de reducir su extensión. En (Molina 2011) han observado que eliminar ciertos segmentos al interior de la oración puede beneficiar algunos métodos de resumen automático.

Con la intención de reunir los datos para la investigación en resumen automático, se creó un sistema que permitió la contribución de cientos de anotadores voluntarios que debían comprimir las oraciones de un texto. Dichos segmentos fueron delimitados previamente. De manera que el sistema mostraba al usuario un texto y él debía elegir si cada segmento debía permanecer o no. Gracias a esta campaña de anotación, se lograron obtener cerca de tres mil resúmenes en tan solo diez semanas y se han logrado analizar diversos aspectos con respecto a la eliminación de segmentos discursivos para la generación de resúmenes en español.

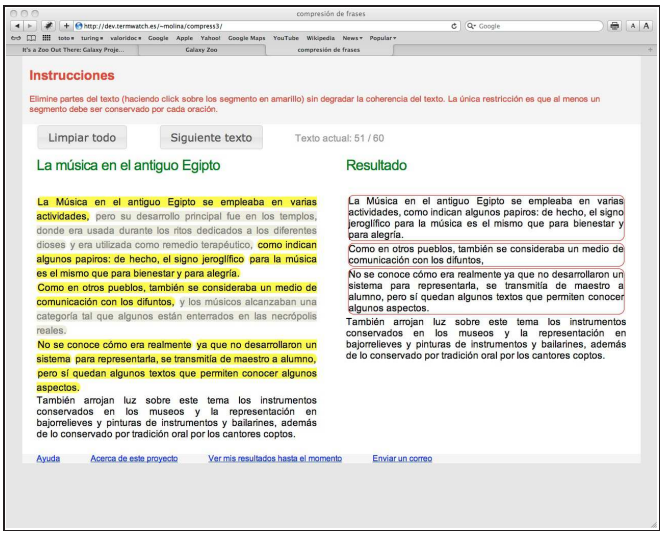

Figura 2. Eliminación de segmentos discursivos para el proyecto de resumen automático

La Web posibilita el desarrollo de proyectos de lingüística computacional que utilizan grandes colecciones de texto y requieren el marcado manual de partes de interés para la investigación

Recopilación de datos para un proyecto de análisis de opinión en microblogs

Cada día millones de individuos publican sus opiniones en la Web usando sitios como Facebook y Twitter. Este último, ha sido objeto de estudios innovadores que buscan evaluar el potencial de la red social en materia de conocimiento de opinión pública. Los resultados descritos en (O'Connor 2010) muestran que existe una alta equivalencia entre las opiniones expresadas en Twitter y los sondeos tradicionales (encuestas en las calles). Por lo tanto, Twitter podría ser una alternativa fiable y económica de futuros sondeos. El proyecto ImagiWeb estudia la imagen de las entidades en la Web a través de las opiniones expresadas en redes sociales[1].

En el marco de este proyecto, se ha desarrollado un sistema de anotación que permitirá reunir rápidamente un corpus con anotaciones acerca de la polaridad de opinión en microblogs. El sistema muestra las publicaciones de la entidad que se quiera analizar; el usuario selecciona un fragmento del texto que expresa una opinión sobre la entidad en cuestión y luego determina si la opinión seleccionada es positiva o negativa. Los datos recopilados servirán para desarrollar métodos para la detección automática de las opiniones, así como su clasificación según la polaridad (opiniones positivas o negativas).

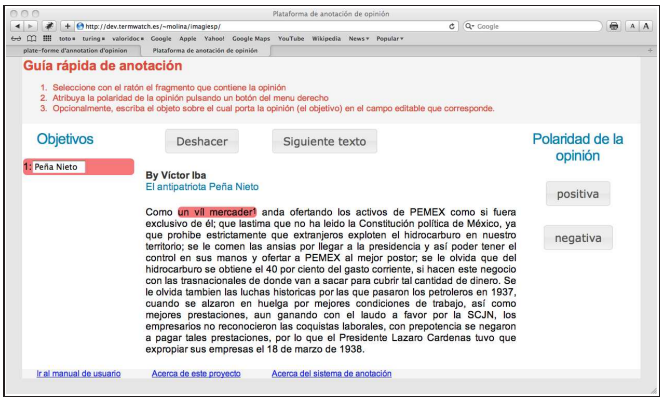

Figura 3. Proyecto de análisis de opiniones en Twitter

[1] dev.termwatch.es/~molina/sentaatool/info/systeme_description.html

 ISSN 2007-0691

Siguiendo normas sencillas en el diseño experimental, los datos obtenidos por no-expertos pueden llegar a ser tan valiosos como los de expertos

Conclusiones

La ciencia ciudadana representa un método viable para la recopilación y el procesamiento de datos en diversos campos científicos. Concretamente, hemos presentado dos proyectos para los cuales elaboramos sistemas de anotación para dos áreas distintas del procesamiento del lenguaje natural. Sin embargo, cabe preguntarse ¿hasta qué punto podemos confiar en los datos obtenidos por voluntarios no-expertos? En (Snow 2008) discuten acerca de esta cuestión y plantean una serie de experimentos para descubrirlo. En su artículo, muestran que los datos de voluntarios no-expertos son casi tan confiables como los de los expertos siempre que el diseño experimental cubra ciertos principios: las descripciones de las tareas, que deben realizar los voluntarios, deben ser tan sucintas como sea posible y la participación requerida debe restringirse a elegir de entre un número limitado de opciones o, si es posible, mediante preguntas de opción múltiple. Tanto en el proyecto de resumen automático como en el de análisis de opinión en microblogs hemos considerado estos principios[2].✳

Agradecimientos. Al Consejo Nacional de Ciencia y Tecnología (México) por la beca doctoral No. 211963 y al proyecto ANR/ImagiWeb (Francia).

REFERENCIAS

1. O'Connor B., Balasubramanyan R., Routledge B.R. y Smith N.A. (2010) "From tweets to polls: Linking text sentiment to public opinion time series". *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pp. 122-129.

2. Fortson L., Masters K., Nichol R., Borne K., Edmondson E., Lintott C., Raddick J., Schawinski K. y Wallin J. (2011) "Galaxy Zoo: Morphological classification and citizen science".

3. Nurmikko T., Dahl J., Gibbins N. y Earl G. (2012) "Citizen science for cuneiform studies".

4. Snow R., O'Connor B., Jurafsky D. y Ng A.Y.(2008) "Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks". *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254-263.

5. Chamberlain J., Poesio M. y Kruschwitz U. (2009) "A new life for a dead parrot: Incentive structures in the Phrase Detectives game". Vol. 9.

6. Evans C., Abrams E., Reitsma R., Roux K., Salmonsen L. y Marra P.P. (2005) "The Neighborhood Nestwatch Program: Participant Outcomes of a Citizen-Science Ecological Research Project", No. 3, Vol. 19, pp. 589-594.

7. Chklovski T. (2003) "Learner: a system for acquiring commonsense knowledge by analogy".*Proceedings of the 2nd international conference on Knowledge capture*, pp. 4-12.

8. Molina A., da Cunha I., Torres-Moreno J.M. y Velazquez-Morales P. (2011) "La compresión de frases: un recurso para la optimización de resumen automático de documentos. No. 3, Vol. 2, pp. 13-27.

SOBRE EL AUTOR

Alejandro Molina es profesor asistente en la universidad de Avignon. Egresado de la maestría en ciencias de la computación de la UNAM y licenciado en computación por la UAM-Iztapalapa. Su trabajo de investigación está centrado en las tecnologías del lenguaje, el reconocimiento de patrones y la lingüística computacional. Es miembro del Grupo de ingeniería lingüística de la UNAM y del equipo de procesamiento del lenguaje en el "laboratoire Informatique d'Avignon" (Francia).

[2]dev.termwatch.es/∼molina/compress4/man

 ISSN 2007-0691