Proyecto FONDEF D09I1185
BNAmericas: Sistema de Descarga y Clasificación de Noticias

Software desarrollado por egresados del DIINF
Emir Muñoz, Rodrigo Campos y Adrián Contreras
Universidad de Santiago de Chile

Problema

Día a día vemos nuevas noticias en los medios de comunicación como periódicos, noticieros, revistas, y otros medios masivos con presencia en Internet. El gran número de estas noticias, y el poco tiempo de las personas para leerlas y procesarlas, impide que las personas se mantengan informadas en todas las áreas de interés. La empresa BNAmericas se dedica a la recolección y edición de noticias financieras relacionadas con Latinoamerica. Gran parte de sus procesos como la recolección y etiquetado, son realizados de forma manual por personas expertas.

Solución

Para ellos se ha desarrollado un sistema automático de **Descarga** y **Clasificación** de noticias, que se divide en dos fases: 1) descarga de páginas HTML de noticias de acuerdo a un conjunto de URL semillas; y 2) un etiquetado o clasificación de las páginas HTML de noticias descargadas, de acuerdo a una taxonomía.

Conceptos básicos

Noticia: redacción de un texto informativo sobre algún hecho novedoso ocurrido dentro de un determinado ámbito específico. Se divide en secciones como título, y cuerpo.
Crawler: también denominado robot o *spider*, es un programa automatizado que se encarga de recolectar e inspeccionar los diferentes contenidos y enlaces entre páginas Web.
URL Semilla: es una URL desde la cual se inicia el procesos de *crawling*.
Etiquetador: clasificador entrenado para asignar etiquetas a documentos en base a un conjunto de características representativas.
Feature: o característica, se extrae a partir del objeto de análisis. Para una noticia, pueden estar presentes en el texto o en su meta-información.
RSS: *Really Simple Syndication*, un formato XML para sindicar o compartir contenido en la Web. Se utiliza para difundir información actualizada frecuentemente a usuarios que se han suscrito a la fuente de contenidos.

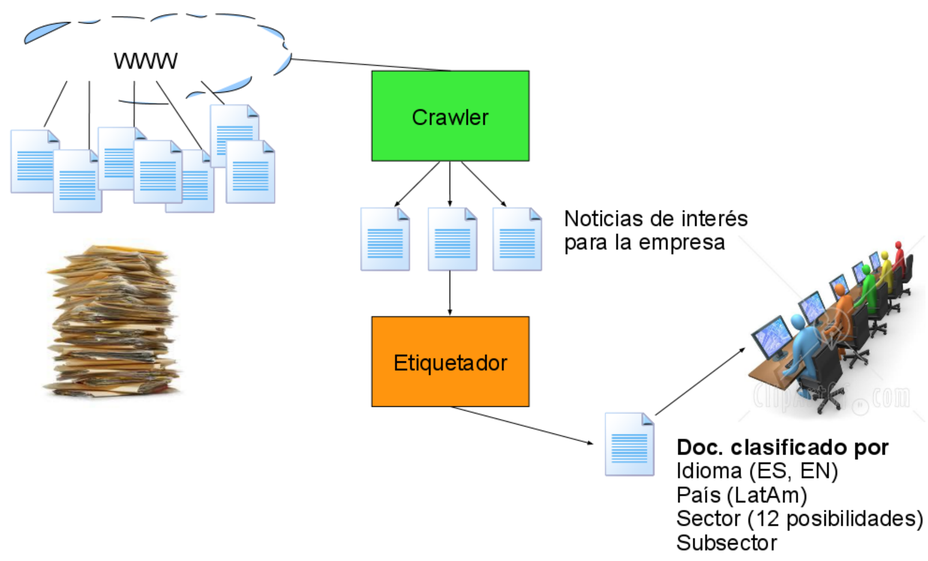
Referencias

[1] C. Manning, P. Raghavan, H. Schtze
Introduction to Information Retrieval
Cambridge University Press, 2008.

[2] G. Adam, C. Bouras, V. Poulopoulos
Efficient extraction of news articles based on RSS crawling
In International Conference on Machine and Web Intelligence, 2010

Diseño de la Solución

El **Descargador** se encarga de descargar las páginas HTML de noticias, según las URL de portales o sitios de noticias. Estas páginas HTML son parseadas, detectando de manera semi-automática su estructura, e identificando las áreas de título y cuerpo de una noticia de interés. Luego, las áreas de una noticia son tomadas por el **Etiquetador**, quién realiza la extracción de *features* definidas, las cuales son pasadas al modelo correspondiente que entrega una etiqueta de acuerdo a la taxonomía manejada.



Métodología Crawler

El **Descargador** o crawler posee un diseño *multi-threading*, así cada *thread* se encarga de una URL semilla (portal). Cada vez que una página HTML es descargada, se identifican todos los *links* en ella y se agregan a la lista de URLs por visitar. Para no redescargar las páginas, se maneja un *buffer* de URL descargadas los últimos n días, y otro de URL visitadas en la actual ejecución.

También, se definen reglas que descartan URLs que no son de interés, por ejemplo aquellas de deporte, cultura, etc. Luego que cada página es descargada, es parseada y procesada para eliminar publicidad, fotos, menús, entre otros, con el fin de dejar sólo el texto de la noticia.

Considerando que las últimas noticias deberían ser las primeras en ser descargadas, se hace uso de los RSS presentes en algunos portales [2]. Del RSS se extraen las URLs de noticias que serán descargadas primero. Una vez extraída la noticia del HTML, ésta es almacenada en un documento XML diferenciando las secciones de la noticia.

Resultados Obtenidos

A partir de una validación utilizando el método *10-fold cross-validation*, utilizando el *corpus* provisto por la empresa, se obtuvieron las siguientes precisiones para los clasificadores:

	EN	ES
Idioma	99,99 %	99,99 %
País	91.18 %	91.42 %
Sector	87.25 %	86.54 %
Subsector	79.42 %	73.11 %

Producto de la descarga de páginas, se obtienen cerca de 800 noticias diferentes diariamente. A las cuales se les aplica el **Etiquetador** con resultados parciales hasta ahora, que se deben evaluar.

Métodología Etiquetador

Los archivos XML son tomados por un **Etiquetador** [1] de noticias, el cual en base a modelos entrenados (usando un *corpus* de 196.725 y 195.519 noticias, en inglés y español respectivamente, provisto por BNAmericas), entrega etiquetas a cada noticia. Estas etiquetas fueron extraídas desde una taxonomía de BNAmericas, y abarcan: **Idioma**, **País**, **Sector económico**, y **Subsector económico**. La clase *Idioma* puede ser: inglés o español. La clase *País* incluye los 7 países más significativos (Brasil, México, Chile, Perú, Argentina, Colombia, Venezuela), en primera instancia, y otros 15 países (Bolivia, Ecuador, Uruguay, entre otros) en segunda instancia. Se consideran 12 posibles etiquetas de la clase *Sector económico*: Minería, Petroleo y Gas, Banca, entre otros; y 36 etiquetas de *Subsectores económicos*.

El Etiquetador, sólo considera los valores de etiquetas con mayor presencia para cada clase en el *corpus*. Esto obedece a un desbalanceo del *corpus*. También se descartaron etiquetas ruidosas.

Conclusiones

El principal problema fue el desbalanceo del *corpus*, teniendo muchas noticias para los países más representativos, y muy pocas para el resto. Similar situación se dio en las otras clases, como Subsectores, donde se tienen 285 etiquetas, pero para algunas de ellas existían sólo 5 noticias o menos. Estos problemas se solucionan acotando las etiquetas a aquellas con suficientes datos.

Los resultados expuestos, se obtienen de una validación cruzada con las noticias etiquetadas del *corpus*. El siguiente paso es validar estos modelos contra otras noticias descargadas directamente de los portales de noticias de diferentes países.