Geolocalización de usuarios en medios sociales mediante análisis de contenidos

Sergio Álvarez Suárez

Tabla de contenidos

Lista de figuras

Capítulo 1. Introducción

1.1. Motivación del proyecto

Aquí la motivación del proyecto

1.2. Alcance

Aquí el alcance del proyecto.

1.3. Estado del arte

El crecimiento exponencial de las redes sociales durante los últimos años ha despertado un gran interés en los diferentes ámbitos de la informática, siendo un claro objetivo comercial para profesionales del sector, así como un nuevo campo de investigación para los investigadores universitarios.

Como consecuencia de todo ello, durante los últimos años han ido apareciendo diversas aplicaciones que, de una u otra manera, se centran en estudiar ciertos aspectos de las redes sociales para poder extraer información acerca de sus usuarios gracias a las diversas publicaciones que estos mismos realizan en sus perfiles.

El estudio de la geolocalización de un usuario a partir de su contenido, sin embargo, es una de las pocas áreas que *tan sólo* agrupa un pequeño número de estudios teóricos, pero en donde no han proliferado herramientas que comprueben de manera empírica los resultados teóricos emitidos por diversos investigadores.

Por ello, en este capítulo se recopilan algunos *papers* que han servido como punto de arranque para este proyecto, así como algunas aplicaciones que comparten características similares.

1.3.1. Papers

Estos son los *papers* más importantes y que han tenido una mayor transcendencia a la hora de desarrollar el proyecto.

Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles

Por Brent Hecht et al. Northwestern University y Palo Alto Research Center

El estudio liderado por Brent Hecht demuestra como aproximadamente el 66% de los usuarios no utiliza el campo de *Localización* en sus perfiles para informar acerca de su localización real. Por tanto, propone como estrategia el conteo de los términos que forman cada tuit con el objetivo de poder realizar cálculos estadísticos que permitan identificar aquellas palabras más indicativas para cada localización.

Esta estrategia será parcialmente utilizada en la implementación del presente proyecto.

Where Is This Tweet From? Inferring Home Locations of Twitter Users

Por Jalal Mahmud et al. IBM Research

Con el objetivo de poder identificar un tuit a nivel de ciudad, este estudio plantea la posibilidad de analizar tres tipos de términos diferentes para localizar una publicación en Twitter:

1. Palabras
2. Hashtags
3. Nombres de lugares (utilizando un *gazetteer* [1] geográfico)

Es interesante observar como empiezan a aparecer pequeñas diferencias entre términos, considerando que en función de su categoría, pueden ofrecer más o menos información geográfica. Esta misma estrategia será también utilizada en el presente proyecto, mediante la extracción de Hashtags, Menciones y N-gramas.

También en este estudio se hace mención a la utilización de un **software de aprendizaje automático**, en este caso WEKA, y su conjunción con un modelo estadístico que realice los cálculos necesarios para el clasificador. El modelo que seleccionaron de manera empírica fue *Naïve Bayes Multimonial*.

TweoLocator: A Non-Intrusive Geographical Locator System for Twitter

Por Yi-Shin Chen et al. National Tsing Hua University

[1] Conjunto de nombres geográficos que, junto con un mapa, constituye una importante referencia sobre lugares y sus nombres

En este estudio, Yi-Shin Chen propone un sistema denominado **TweoLocator** que funciona como un *framework* el cual, a través de diferentes etapas, asegura ofrecer unos resultados altamente fiables acerca de la localización de un usuario en Twitter.

Este paper ofrece el interesante concepto de n-gramas, que será utilizado en el presente proyecto como una de las estrategias para la detección de términos con posibilidad de ofrecer contenido geolocalizable.

A Multi-Indicator Approach for Geolocalization of Tweets

Por Axel Schulz et al. SAP Research

Presenta un sistema muy interesante mediante la utilización de formas poligonales en 3D para decidir la localización de un tuit y usuario. Cada polígono tiene dos valores de calidad en base al indicador que los define. Los polígonos se superponen, y la intersección de mayor altura es el área con más probabilidades de contener el tuit analizado.

Para obtener la información de cada tuit, utiliza varios sistemas como la **DBPedia** (http://dbpedia.org/) o **Foursquare** (https://es.foursquare.com/) para reconocer entidades y topónimos. Su dependencia de sistemas externos impide que sea capaz de deducir una localización a través del contexto si esta no contiene ninguna referencia a una entidad localizable.

Inferring the Origin Locations of Tweets with Quantitative Confidence

Por Reid Priedhorsky et al. Los Alamos National Laboratory y Northeastern Illinois University

Este nuevo estudio presenta otra estrategia para encontrar términos fuertemente localizados en base al uso de *n-gramas*. Para ello, los investigadores se centraron en extraer bigramas de los siguientes campos:

- Campo de Localización
- Contenido del tuit
- Zona horaria
- Idioma seleccionado en el perfil del usuario

A su vez, desarrollaron un modelo estadístico propio, basado en un clasificador Gaussiano, el cual aplicaban sobre aquellos bigramas que superaban un número mínimo de apariciones.

Home Location Identification of Twitter Users

Por Jalal Mahmud et al. IBM Research

Otros *papers* de interés

- **You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users** *por Zhiyuan Cheng et al. Texas A&M University*
- **Location Type Classification Using Tweet Content** *por Haibin Liu et al. The Pennsylvania State University*
- **TweetLocalize: Inferring Author Location in Social Media** *por Evan Sparks et al. University of California-Berkeley*
- **Inferring the Location of Twitter Messages Based on User Relationships** *por Clodoveu A. Davis Jr. et al. Universidade Federal de Minas Gerais*
- **Geolocation Prediction in Social Media Data by Finding Location Indicative Words** *por HAN Bo et al. University of Melbourne*

1.3.2. Aplicaciones web

A continuación, se recopilan algunas aplicaciones con características similares a las del presente proyecto, y otras que, sin centrarse en el estudio de la geolocalización del contenido de manera específica, utilizan **Twitter** como fuente de información y realizan un estudio a gran escala sobre las publicaciones realizadas.

Trendsmap

Trendsmap (http://trendsmap.com/) es una aplicación web que muestra en tiempo real las tendencias en Twitter para cada localización a escala mundial. Según datos ofrecidos por la propia web en su página de *FAQ*, gestionan un volumen de tuits del orden de **80 millones al día**.

Su objetivo principal consiste, por tanto, en localizar las tendencias específicas de cada localización, sin ofrecer ninguna información acerca de las posibilidades de que uno u otro tuit que contengan dicha tendencia puedan pertenecer a una localización en concreto.

No se ha podido encontrar ninguna información acerca del tipo de algoritmo que utiliza Trendsmap para determinar la localización de un tuit (en pos de poder determinar la tendencia para un alto número de los mismos).

What the Trend

What the Trend (http://whatthetrend.com/faq) se centra en ofrecer al usuario una explicación acerca de los propios *Trending Topics* identificados por Twitter para cada localidad.

En este caso, la aplicación no incluye ningún tipo de algoritmo para adivinar la localidad de un volumen de tuits, si no que únicamente recoge las tendencias previamente analizadas y localizadas por Twitter.

Klout

Klout (http://klout.com/home) se describe como un servicio capaz de obtener la influencia de un usuario en la red a través de sus publicaciones y relaciones en redes sociales. Durante sus primeros años fue objetivo de varias inversiones millonarias que sacaron a la luz la gran importancia que tiene a nivel empresaria el análisis de los grandes volúmenes de información que se generan a diario en las redes sociales por parte de los propios usuarios.

1.4. Aspectos teóricos

Aquí los aspectos teóricos.

Capítulo 2. Planificación y gestión del proyecto

Capítulo 3. Análisis del sistema

Capítulo 4. Desarrollo e implementación del sistema

4.1. Obtención de datos

El primer paso en el desarrollo del proyecto, fue la creación de un sistema capaz de recolectar tuits con el objetivo de obtener material de entrenamiento sobre el que aplicar las diferentes estrategias planteadas.

Este sistema debía ser parametrizable, con el objetivo de poder configurar en cada ejecución el tipo de tuits que se querían obtener. En base a estos requisitos, se utilizó la **API Streaming de Twitter** (https://dev.twitter.com/docs/api/streaming), la cual aporta dos características principales:

- Capacidad de obtener datos en tiempo real de Twitter de manera ininterrumpida
- Capacidad de establecer filtros sobre el flujo de entrada:
  - Filtro por idioma del tuit
  - Filtro por localización del tuit (mediante el uso de *bounding boxes*)

Para realizar la conexión entre el sistema desarrollado y la API de Twitter se utilizó la biblioteca **Twitter4j** (http://twitter4j.org/en/index.html), la cual aunque originalmente está desarrollada para ser utilizada sobre Java, es perfectamente utilizable en Scala gracias a la compatibilidad entre ambos lenguajes mediante la Java Virtual Machine. La ventajas de utilizar una biblioteca construida sobre la API original de Twitter es que algunos de los problemas más habituales se solucionan a través de nuevas capas de abstracción:

- Autenticación OAuth2 simplificada mediante clases propias de la biblioteca
- La API de Twitter4J para utilizar el Streaming de Twitter permite aislar al desarrollador de la necesidad de mantener activa la comunicación HTTP manualmente para estar conectado al Streaming de Twitter.

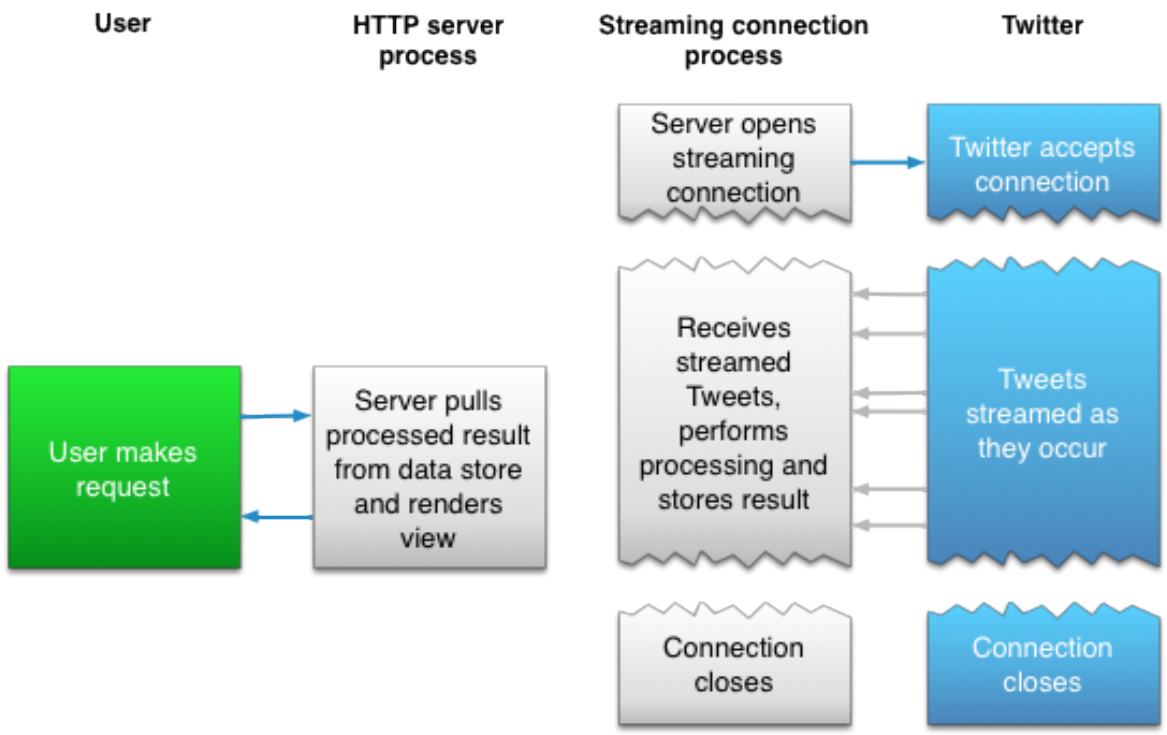

Figura 4.1. Modelo de comunicación entre un cliente y la API Streaming de Twitter (https://dev.twitter.com/docs/api/streaming)

4.1.1. Almacenamiento de datos

Uno de los puntos más importantes que planteó el sistema para recolectar tuits era en qué formato sería más adecuado serializar los datos obtenidos.

En un primer momento se barajó la posibilidad de utilizar el formato CSV, el cual permitiría acceder de manera rápida al número de tuits guardados y realizar operaciones sencillas en línea de comandos mediante operaciones `grep`. Esta decisión fue cancelada al realizar los primeros experimentos y comprobar como el guardado de ciertos datos en formato CSV presenta muchas dificultades para poder solventar todos los casos esquina que se presentan con la aparición de contenido complejo que pueda incluir comas, comillas y otros signos de puntuación (aún en el caso de utilizar bibliotecas especializadas como OpenCSV - http://opencsv.sourceforge.net/) combinados con caracteres extraños como Emoji (http://www.unicode.org/faq/emoji_dingbats.html).

Como consecuencia de los resultados anteriores, y apoyado en el soporte nativo ofrecido por Scala, se utilizó XML como el lenguaje de marcado que mejor podría serializar y estructurar los datos obtenidos a través de Twitter4j. El siguiente fragmento

de código permite ver lo sencillo que es serializar un objeto en Scala a XML mediante la utilización de literales:

```
class Tweet(id:String, username: String, name:String, location: String,
 timezone: String, createdAt:String, latitude: String,
            longitude: String, text: String) {
  def toXML =
    <tweet>
      <id>
        {id}
      </id>
      <username>
        {username}
      </username>
      <name>
        {name}
      </name>
      <location>
        {location}
      </location>
      <timezone>
        {timezone}
      </timezone>
      <createdAt>
        {createdAt}
      </createdAt>
      <latitude>
        {latitude}
      </latitude>
      <longitude>
        {longitude}
      </longitude>
      <text>
        {text}
      </text>
    </tweet>
}
```

4.1.2. Parámetros del sistema

Como parte de los requisitos del sistema, era necesario ofrecer la capacidad de parametrizar la ejecución para poder obtener un tipo de resultados u otros. A continuación se muestra un ejemplo del fichero **`properties`** que se ha utilizado para indicar al sistema algunos parámetros:

- **time_in** y **time_to_collect**: permiten establecer al recolector un tiempo de ejecución representado en diferentes magnitudes.
- **file_name**: nombre del fichero de salida.
- **coordiantes_mandatory**: **boolean** que indica si los tuits recolectados deben contener o no información geográfica adjunta.
- **filter_language**: idioma en el que se desean obtener los tuits.
- **stop_words_file**: debido a restricciones de Twitter, es necesario proveer una lista de términos cuando se intenta realizar un filtrado por idioma. Con el objetivo de restringir lo mínimo posible el número de tuits a obtener, se provee una lista de *stop words* del idioma por el que se esté filtrando.
- **filter_bounding_boxes_file**: fichero **XML** que contiene los *bounding boxes* sobre los que se realizará el filtrado para aquellas ejecuciones que requieran tuits localizados en un área en concreto.

El siguiente fragmento de código muestra un fichero donde se aplican a la vez todas las propiedades. En un caso real, sólo se aplicarían aquellas que tuvieran sentido para el resultado que se quisiera obtener. Por tanto, no tendría sentido que estuvieran habilitadas a la vez las propiedades **filter_language** y **filter_bounding_boxes_file**

```
# Valid values for time_in property:
#
# DAYS
# HOURS
# MICROSECONDS
# MILLISECONDS
# MINUTES
# NANOSECONDS
# SECONDS
time_in=DAYS
time_to_collect=1

# File to save the results
file_name=ES-Spain.xml

coordinates_mandatory=true
filter_language=es
stop_words_file=spanish-stop-words.txt
filter_bounding_boxes_file=es_bounding_boxes.xml
```

4.1.3. Ejemplo de resultados

Un ejemplo de los resultados obtenidos por el recolector sería el siguiente:

```
<tweets>
  <tweet>
    <username>
      gaabriforner
    </username>
    <location>
      Málaga
    </location>
    <timezone>
      Athens
    </timezone>
    <createdAt>
      2014-03-04 21:53
    </createdAt>
    <latitude>
      -4.437747
    </latitude>
    <longitude>
      36.7055494
    </longitude>
    <text>
      y ante todo a echarle fuerza d voluntad y ganas para conseguir lo
 que quiero!!
    </text>
  </tweet>
</tweets>
```

4.2. Análisis de datos

4.3. Entrenamiento del modelo de análisis de datos