Graph Visualization Tool for Twittersphere users based on a high-scalable Extract, Transform and Load System

Pablo Aragón, Íñigo García and Antonio García

May, 27th 2011
## INDEX

### INTRODUCTION
- Cierzo Development and SMMART
- Structure of Twitter
- Volume of Twitter
- Detection of influencers

### DISTRIBUTED COMPUTATION
- Hadoop
- Amazon EC2

### PIPELINE DESIGN
- Crawling Module
- Metadata Extraction Module
- Indexing Module
- Graph Visualization Module

### RESULTS
- Western Sahara Conflict
- Patxi López
- Conclusions
- Future work
### INTRODUCTION: CIERZO DEVELOPMENT AND SMMART

SMMART (Social Media Marketing Analysis and Reporting Tool) is the system developed by Cierzo Development for:

- Corporate social reputation
- Measuring effectiveness of marketing campaigns
- Detection of new trends
INTRODUCTION: STRUCTURE OF TWITTER

Structure of a profile
INTRODUCTION: STRUCTURE OF TWITTER

A user can set a relationship with another user by:

- **Reply**: Update that begins with @username
- **Mention**: Update that contains @username in the body of the tweet
- **Retweet**: Update that contains the body of another user tweet by specifying the original author
INTRODUCTION: VOLUME OF THE TWITTER

More than 200M users publishing millions of tweets per day
## INTRODUCTION: DETECTION OF INFLUENCERS

<table>
<thead>
<tr>
<th>Rank</th>
<th>User</th>
<th>Link</th>
<th>Followers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Lady Gaga</td>
<td><a href="http://www.ladygaga.com">http://www.ladygaga.com</a></td>
<td>9.814.456</td>
</tr>
<tr>
<td>2</td>
<td>Justin Bieber</td>
<td><a href="http://www.youtube.com/justinb">http://www.youtube.com/justinb</a></td>
<td>9.467.962</td>
</tr>
<tr>
<td>3</td>
<td>Barack Obama</td>
<td><a href="http://www.barackobama.com">http://www.barackobama.com</a></td>
<td>7.862.926</td>
</tr>
<tr>
<td>4</td>
<td>Britney Spears</td>
<td><a href="http://www.britneyspears.com">http://www.britneyspears.com</a></td>
<td>7.717.271</td>
</tr>
<tr>
<td>5</td>
<td>Kim Kardashian</td>
<td><a href="http://kimkardashian.celebuzz">http://kimkardashian.celebuzz</a></td>
<td>7.403.408</td>
</tr>
<tr>
<td>6</td>
<td>Katy Perry</td>
<td><a href="http://www.katyperry.com">http://www.katyperry.com</a></td>
<td>7.149.680</td>
</tr>
<tr>
<td>7</td>
<td>Ashton Kutcher</td>
<td><a href="http://www.facebook.com/Ashton">http://www.facebook.com/Ashton</a></td>
<td>6.633.714</td>
</tr>
<tr>
<td>8</td>
<td>Ellen DeGeneres</td>
<td><a href="http://www.ellentv.com">http://www.ellentv.com</a></td>
<td>6.577.668</td>
</tr>
<tr>
<td>9</td>
<td>Taylor Swift13</td>
<td><a href="http://twitter.com/taylorswift">http://twitter.com/taylorswift</a></td>
<td>6.204.923</td>
</tr>
<tr>
<td>10</td>
<td>Shakira</td>
<td><a href="http://www.shakira.com">http://www.shakira.com</a></td>
<td>5.858.885</td>
</tr>
</tbody>
</table>

Old metrics based on data as:

- **Absolute info:** Number of followers
- **Relative info:** Quotient of following users and followers
Available search engines track Twitter and list results, but they do not set a value to the users from the response.
Spanish voters head to the polls, as city square protests continue

Millions turn out to elect municipal councils and regional governments, despite protests over politics-as-usual.

Gilles Tremlett in Madrid
guardian.co.uk, Sunday 22 May 2011 19:46 BST
Article history

Protesters react as they vote in regional elections.

Opinion

'Yes we camp' activists hit Spanish streets

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.

Leila Nachawati
Leila Nachawati, 22 May 2011 10:34

When Spanish officials banned a protest camp in central Madrid ahead of elections, thousands defied the ruling.
Management of large volumes at the lowest cost

Automatic adjustment to the daily growth of users and the oscillations in the frequency of publication
INTRODUCTION
DISTRIBUTED COMPUTATION
PIPELINE DESIGN
RESULTS

DISTRIBUTED COMPUTATION: HADOOP

Map Reduce

Distributed File System
Definition of a Hadoop node as a machine image in Amazon Elastic Compute Cloud.

The system balancing mechanism adds and removes Hadoop nodes in real time on demand.
PIPELINE DESIGN

- INTRODUCTION
- DISTRIBUTED COMPUTATION
- PIPELINE DESIGN
- RESULTS

CRAWLING MODULE
METADATA EXTRACTION MODULE
INDEXING MODULE
GRAPH VISUALIZATION MODULE

PIPELINE DESIGN Diagram:

- Twitter
  - URLs
  - Crawling Module
  - HTML
  - Metadata Extraction Module
  - Metadata sets
  - Indexing Module
  - Documents

- Web User Interface
  - Query
  - Index

- Graph Visualization Module
  - Graph
  - Documents
Based on Nutch

1. Crawl the Twitter profiles stored in a DB
2. Extract outlinks to new profiles
The portion of HTML of a tweet contains a set of metadata:

- Textual content
- Publication date
- Author
- Mention to other users
PIPELINE DESIGN: INDEXING MODULE

Apache Solr (enterprise search server based on Lucene)

- Sorting algorithms
- Stemming
- Stopwords filters
- Faceted searches

Multicore architecture sharding by publication date.
The Graph Visualization module transforms the responses from the index into a graph by the force-based multilevel layout Yifan Hu’s algorithm provided in Gephi Toolkit.
In November 2010, Moroccan security forces involved in a camp in Western Sahara. This action was criticized by part of the Spanish society.
### RESULTS: WESTERN SAHARA CONFLICT

**Search**
- content: ‘sahara’
- language: ‘es’
- date: [2010-11-10 TO 2010-11-18]

**Results**
- 1721 users
- 3925 tweets
- 707 mentions
RESULTS: WESTERN SAHARA CONFLICT

Pepelu Saliquet
@Pepeluibiza iPhone: 38.966303, 1.497935
News, politic, history, technology, nature & Music, in love with my home Ibiza, our nature & sea. I always enjoy sweet, good & groovy human vibes.
http://wefollow.com/Pepeluibiza

pituskaya
@pituskaya Málaga, España
Otro mundo es posible. Derechos Humanos para todos.
http://www.tesladream.org/activismo/activismoenred.html

Free Sahara
@freedomsohara Sahara Occidental
Free Sahara es un grito por la dignidad y el respeto a un pueblo perseguido, torturado y olvidado por la comunidad internacional. #freesahara
Patxi López holds the position of the President of the Basque Country Government. His campaign included strategies in social networks.
RESULTS: PATXI LÓPEZ

Search

- mention:'patxi_lopez'
- language:'es'
- date:[2010-11-10 TO 2010-11-18]

Results

- 186 users
- 196 tweets
- 366 mentions
RESULTS: PATXI LÓPEZ

Patxi Lazcoz
@patxi_lazcoz  Vitoria-Gasteiz
Twitter del alcalde de Vitoria-Gasteiz.
http://blogs.vitoria-gasteiz.org/alcalde/

PSOE La Rioja
@PSOE_LaRioja  La Rioja
Partido Socialista de La Rioja (Twitter oficial)
http://www.psoe-larioja.es

Javier Remírez
@javierremirez  iPhone: 42 833614, -1 562662
Ciudadano navarro, consciente y ciberactivista.
Picapetitos a tiempo parcial, Jefe de Personal en horas pares...oh! y socialista total (uy! lo k he dicho)
http://elpurgatorioexiste.wordpress.com
The implemented tool identifies main influencers in a specific topic or around a concrete user.

- The high-scalable design adapts to a large social network as Twitter.
- Enterprises can deploy social media monitoring systems using exclusively open source technologies.
- The tool provides information for performing crisis management.
RESISTENCIA A LA DICTADURA ECONÓMICA CIENTÍFICA
ERROR: DEMOCRACY NOT FOUND

<table>
<thead>
<tr>
<th>Tag</th>
<th>Evolución</th>
</tr>
</thead>
<tbody>
<tr>
<td>SYMES</td>
<td>13</td>
</tr>
<tr>
<td>estabilidad</td>
<td>10</td>
</tr>
<tr>
<td>QSS</td>
<td>4.9</td>
</tr>
<tr>
<td>UNISONO</td>
<td>4.7</td>
</tr>
<tr>
<td>COHEREH*</td>
<td>2.1</td>
</tr>
<tr>
<td>ERROR</td>
<td>2</td>
</tr>
<tr>
<td>APLAUD*</td>
<td>0.4</td>
</tr>
<tr>
<td>SERTEL</td>
<td>0.3</td>
</tr>
</tbody>
</table>

[Diagrama de red con etiquetas]

[Gráfico de barras con datos]
RESULTS: FUTURE WORK

- New versions for more social media sources
- Real-time results
- New data mining applications
  - Predictive models
Thanks for your attention