II Seminario E-Lectra

Lines of work in Web Mining, knowledge extraction and social network analysis

Carlos García Figuerola, José Luís Alonso Berrocal, Ángel Zazo Rodríguez

Institute of Science and Technology Studies
University of Salamanca
{figue, berrocal, angelzazo}@usal.es
Editing content in a collaborative environment: The case of the Spanish Wikipedia

Ángel Zazo Rodríguez
angelzazo@usal.es
Institute of Science and Technology Studies
University of Salamanca
Introducción

Indicators of scientific and technological culture ... Wikipedia

- 36 M page views a day in 2013 (only articles)
  Alexa Spain: top 8
- Volunteers editing
- Collaborative editing
- Quality but errors and vandalism
  - Politics, neutral point of view, content guidelines, editing guidelines...

Ángel Zazo Rodríguez. Editing content in a collaborative environment: The case of the Spanish Wikipedia
Introducción

Content management system: data base storage

- Pages
  - Articles and redirections
- Categories
- Editors
- Talk pages
- Users talk pages
- Interwiki links
- Revision history
- Users actions
Introducción

- Complex structure of data base

- Big data
  - ~ 4,5M of pages:
    - ~ 1,1M articles
    - ~ 1,5M redirections
  - ~ 240K categories
  - ~ 63M interwiki links
  - ~ 3M registered users
  - ~ 5M anonymous users
  - ~ 67M editions
  - ~ 36M page views a day (only articles, 2013)
Introducción

Data source:

- Specific data in real time*:
  - Wikimedia Toolserver – Wikimedia Tool Labs
- Dump data every two weeks
  - Wikimedia Dumps
    - Date: 01/11/2013 [~1,5TB]
    - Access in 2013 [~1,3TB]
Statistics (v. 2013-11-01)

- Number of pages: 4,377,848
- Number of articles (no redirects): 1,027,168
- Number of edits (articles): 49,161,326
- Number of categories (only articles): 197,343
- Number of links (between articles): 30,023,103
- Number of editors: 5,685,536
- Number of article views in 2013: $13 \times 10^9$
Editing: different kinds of editors

<table>
<thead>
<tr>
<th>Editors</th>
<th>#</th>
<th>New Articles</th>
<th>Editions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registered (no admin)</td>
<td>467,199</td>
<td>388,743</td>
<td>13,904,570</td>
</tr>
<tr>
<td>Admin (sysop, bureaucrat, rollbacker, checkuser...)</td>
<td>784</td>
<td>492,964</td>
<td>9,505,670</td>
</tr>
<tr>
<td>Bots</td>
<td>366</td>
<td>3,832</td>
<td>10,559,405</td>
</tr>
<tr>
<td>Unregistered</td>
<td>5,217,187</td>
<td>141,629</td>
<td>15,191,681</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>5,685,536</strong></td>
<td><strong>1,027,168</strong></td>
<td><strong>49,161,326</strong></td>
</tr>
</tbody>
</table>

Number of editors (UnREG excluded)

Ángel Zazo Rodríguez. *Editing content in a collaborative environment: The case of the Spanish Wikipedia*
**Statistics (v. 2013-11-01)**

- **Distribution of categories per article**
- **Distribution of outlinks per article**
- **Distribution of inlinks per article**
- **Distribution of article length (bytes)**

- **Charles_Darwin** is assigned to 48 categories
- **Historia_del_Arte** has 4,237 outlinks
- **Estados_Unidos** has 184,056 inlinks
Editorial habits

Ángel Zazo Rodríguez. *Editing content in a collaborative environment: The case of the Spanish Wikipedia*
Editorial habits

Automated conversion

Changing Interwiki links to Wikidata

Approximately 115,000 articles per month are revised (bots excluded)

Ángel Zazo Rodríguez. Editing content in a collaborative environment: The case of the Spanish Wikipedia
Editorial habits

<table>
<thead>
<tr>
<th>Number of different editors (no bots)</th>
<th>Number of articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>5,000</td>
</tr>
<tr>
<td>10</td>
<td>10,000</td>
</tr>
<tr>
<td>100</td>
<td>100,000</td>
</tr>
<tr>
<td>1,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td>10,000</td>
<td>10,000,000</td>
</tr>
</tbody>
</table>

**Distribution of number of articles edited by human editors (no bots)**

- Dragon Ball has been edited by 3,147 different human editors
- Rosarinagazo has edited 54,836 different articles

**Distribution of number of human editors by number of edited articles**

Ángel Zazo Rodríguez. *Editing content in a collaborative environment: The case of the Spanish Wikipedia*
Editorial habits

Active days per editor (%)

Average edits per day

Ángel Zazo Rodríguez. Editing content in a collaborative environment: The case of the Spanish Wikipedia
Editor origin

Origin of editors

Origin of editors for the Spanish Wikipedia.


- Editors (5,223,952)
- New articles (624,002)
- Edits (25,436,407)

Ángel Zazo Rodríguez. Editing content in a collaborative environment: The case of the Spanish Wikipedia.
Article view statistics (year 2013)

Number of page views in 2013 (only articles): $13 \times 10^9$

Most viewed articles:

<table>
<thead>
<tr>
<th>January</th>
<th>February</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>Jun</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reyes_Magos (1799482)</td>
<td>Día_de_San_Valentín (1671271)</td>
<td>Día_Internacional_de_la_Mujer (3268741)</td>
<td>Ella_Fitzgerald (1370257)</td>
<td>Día_de_la_Madre (1530565)</td>
<td>Día_del_padre (1493260)</td>
</tr>
<tr>
<td>Facebook (1008564)</td>
<td>Nicolás_Copérnico (1068285)</td>
<td>Francisco_(papa) (2660326)</td>
<td>Arroba_(símbolo) (1065131)</td>
<td>Arroba_(símbolo) (1091357)</td>
<td>Arroba_(símbolo) (1053283)</td>
</tr>
<tr>
<td>Arroba_(símbolo) (864046)</td>
<td>Facebook (1019703)</td>
<td>Harlem_Shake_(meme) (1872583)</td>
<td>Día_de_la_Tierra (975260)</td>
<td>Facebook (882493)</td>
<td>Facebook (960558)</td>
</tr>
<tr>
<td>One_Direction (853703)</td>
<td>Go (908219)</td>
<td>Hugo_Chávez (1652315)</td>
<td>Facebook (889354)</td>
<td>Sistema_Solar (698119)</td>
<td>Emma_Stone (846443)</td>
</tr>
<tr>
<td>Go (839668)</td>
<td>Bandera_de_México (907505)</td>
<td>Semana_Santa (1012442)</td>
<td>Hotmail (783077)</td>
<td>Baloncesto (645772)</td>
<td>Copa_FIFA_Confederaciones_2013 (833945)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>July</th>
<th>August</th>
<th>September</th>
<th>October</th>
<th>November</th>
<th>December</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook (1127149)</td>
<td>Iusacell (1804695)</td>
<td>Tyrrell_Racing (1114923)</td>
<td>Halloween (1704629)</td>
<td>Hermann_Rorschach (2026857)</td>
<td>Paul_Walker (2462939)</td>
</tr>
<tr>
<td>Arroba_(símbolo) (1053844)</td>
<td>Estadio_Corona (1162021)</td>
<td>Independencia_de_México (927160)</td>
<td>Go (1128943)</td>
<td>Doctor_Who (1784042)</td>
<td>Carlos_Finan (1971717)</td>
</tr>
<tr>
<td>Despicable_Me (846942)</td>
<td>Arroba_(símbolo) (849684)</td>
<td>Go (730248)</td>
<td>André_Jacques_Garnerin (1088985)</td>
<td>Go (997608)</td>
<td>Nelson_Mandela (1918722)</td>
</tr>
<tr>
<td>Go (782633)</td>
<td>Go (807873)</td>
<td>Emma_Stone (657659)</td>
<td>Día_de_Muertos (1006279)</td>
<td>Carlos_Fuentes (980378)</td>
<td>Santoral_catolico (1638522)</td>
</tr>
<tr>
<td>Franz_Kafka (759381)</td>
<td>José_de_San_Martín (770620)</td>
<td>Julianne_Moore (631616)</td>
<td>Arroba_(símbolo) (1005193)</td>
<td>Arroba_(símbolo) (933632)</td>
<td>Navidad (1039119)</td>
</tr>
</tbody>
</table>
Lines of work
Lines of work

Science & Technology in Wikipedia

- Improvements detecting S&T articles
- Applying statistical and other studies to S&T articles

Ángel Zazo Rodríguez. *Editing content in a collaborative environment: The case of the Spanish Wikipedia*
Lines of work

- Editions & Page views
  - Life events
  - Natural disasters

- Vandalism
  - Addition, removal, or change of content, in a deliberate attempt to compromise the integrity of Wikipedia
  - Detect vandalism
    - Reverting edits... but reverts can be done for every editor
    - Use the `sha1sum` field of “revision” table
    - Use the information of `Talk pages`
Lines of work

Conflicts and edit wars

- Different causes
- Related to
  - Life events
  - Natural disasters

Detect conflicts:

- Detect *reverts*
- Using *Talk pages* (length)
- Some heuristics can be applied


Ángel Zazo Rodríguez. *Editing content in a collaborative environment: The case of the Spanish Wikipedia*
Editing content in a collaborative environment: 
The case of the Spanish Wikipedia

Ángel Zazo Rodríguez
angelzazo@usal.es
Institute of Science and Technology Studies
University of Salamanca