Réseau des bibliothèques de Suisse occidentale
Westschweizer Bibliotheksverbund
Rete delle biblioteche della Svizzera occidentale
Library Network of Western Switzerland

The New RERO Statistics Services

Invenio User Group Workshop 2015

Johnny Mariéthoz

2015/10/05

Introduction I

- institutions need statistics for reports and analysis
- as instance manager we need statistics to understand user behavior

Introduction II

- a statistics module exists in Invenio
  - unfortunately it includes bots accesses (widely more frequent than humans)
  - it does not give enough detailed information
- awstats
  - too general
  - poor robot detection
  - inadequate user interface (lacks of multiple filters)
- google analytics
  - difficult to add instance specific information such as collections
  - data protection (privacy)

we therefore decided to create our own module

Workflow

1. parse apache log to retrieve access information
   - IP address
   - datetime
   - URL
   - http status
   - referrer
2. filters accesses to keep only useful accesses (to avoid bots trafic)
3. on the fly Elasticsearch indexing
4. users can access the statistics using a specific web page
   - angular based interface using ajax request to query the ES index

Technology

- indexing
  - apache logger that pipe into a home-made indexing script
  - on the fly indexing using Invenio API to enrich data
  - black IPs creation
  - specific Elasticsearch index
  - use UASparser: User Agent String parser
  - GeoLiteCity: geo-localisation IP based database
- frontend
  - based on AngularJS
  - use ajax request to a Flask-API

Apache Log Parsing

- on the fly indexing using a script CustomLog "| log_indexing.py"
- regular expression to parse logs
- remove non pertinent data (images, css, etc.)
- remove bots accesses

Enrich the Data

- add informations using Invenio API, UASparser and GeoLiteCity
  - with a *recid* and Invenio API we add:
    - insitution
    - document type
    - UDC domains
    - patrimonial collection
- based on the URL we add:
  - *recid*
  - file name
  - request type: simple search, advanced search, file uploads, etc.
  - language of the interface
  - referrer
- using GeoLiteCity and the IP address we add
  - country

Frontend

- use AngularJS with bootstrap and Highcharts
- ajax based Elasticsearch queries through a JSON-RPC Flask server
- has been tested by several institutions and the feedback is widely positive

Considered as Human Access

- corresponds to a predefined list of valid URLs (search, record, file, etc.)
- referrer does not match Bots regular expression
- IP in white list tips
- IP not in black list

White List

- RERO Institutions IPs extracted from Invenio (web access)

Black List IPs Rules

Used to remove undeclared bot accesses.

- more than 1 access to "/robots.txt"
- more than 20 accesses to MarcXml (/export/xm)
- more than 20 accesses to MODS (/export/xo)
- more than 20 accesses to subformat (subformat*)
- more than 20 accesses to file version (?version=1)
- more than accesses to records or files with an unlikely "referrer" (direct access from the main page):
  - http://doc.rero.ch
  - https://doc.rero.ch
  - http://doc.reroc.ch/
  - https://doc.reroc.ch/

Demo

Let's see how it looks like:

- `http://doc.rero.ch/stats/#?ln=en`
- `http://doc.rero.ch/stats/#?ln=en&cc=PRESS`

Learnt from the Data

- fun to play and learn using the new interface
- Google is our friend (of course), Bing is not (sorry)!
- bug detected for Scholar since our migration (will be fully re-indexed in late-January)
- save time to produce statistics (previously needed manual work)
- Multivio (home-made PDF reader) is used but not always
- most of accesses are files
- Invenio search interface is still used (not only Google)

And Next?

- display statistics in the Invenio detailed record view
- CSV export for reports usage

References

- Elasticsearch `https://www.elastic.co/fr/`
- Flask Python Web Framework `http://flask.pocoo.org/`
- JSON-RPC Flask package
  `https://github.com/cenobites/flask-jsonrpc`
- Highcharts Javascript Library
  `http://www.highcharts.com/`
- HTML/CSS Framework `http://getbootstrap.com/`
- IP geo localisation database `http://dev.maxmind.com/`
- User Agent Database
  `https://github.com/chetan/UASparser`
- Invenio Software `http://invenio-software.org/`