Wikipedia data analysis for researchers

•Keyboard_light.jpg; autor: PJ (Wikimedia Commons)

Felipe Ortega
felipe.ortega@urjc.es
Identi.ca/twitter: @jfelipe

WPAC-2012 (Berlin). June 29, 2012.

Summary

1) Preparing for Wikipedia data analysis (75').
   - Understanding Wikipedia data sources.
   - Data preparation and storage.
   - Available (FLOSS) tools.
2) Conducting Wikipedia data analysis (90').
   - Methodology and (FLOSS) tools.
   - Example cases.
     - General statistics.
     - Study of inequalities.
     - Logging actions.

1. Preparing for Wikipedia data analysis

1.1 Understanding data sources

- Activity vs. Traffic.
  - Activity: edits, new registrations, blocks...
  - Traffic: browsing requests (read, edit, preview, save, search).
- We will focus on activity data sources.
  - In particular, on Wikipedia dump files.
- For traffic:
  - Counting page views.
  - Traffic statistics.

1.1 Understanding data sources

- Obvious choice: web scrapping.
    - Not recommendable.
    - It can generate too much traffic.
    - Risk of getting banned.
    - Extra work to interpret data and filter out format.
- It does not worth the time and effort...
    - Except for extremely well-justified cases.

1.1 Understanding data sources

- MediaWiki API.
  - For reading and writing (with user account and correct privileges).
  - Read available documentation first.
  - Multiple output formats: JSON, XML, YAML...
- Page for the MediaWiki API.
- API doc in the English Wikipedia.

1.1 Understanding data sources

- The toolserver(s).
  - Explained in a parallel workshop.
  - Contains mirrors of all databases for all Wikimedia projects.
  - Good environment for testing applications and accessing “live“ data.
  - Shared machine, observe etiquette rules and use resources with care.
- Revision history statistics.
- User edits.

1.1 Understanding data sources

- Wikipedia dump files (our focus).
  - Snapshot, some delay for huge languages.
  - Complete freedom to operate locally with your data (burn your machine!!).
  - Opportunities for (pre)computing additional metadata (more on this later).
- Download center.
- Data dumps.
- Available info and formats.

1.1 Understanding data sources

- Dump files.
  - Stub-meta-history.
  - Pages-meta-history.
  - Pages-meta-current.
  - Page links, external links, interlanguage links.
  - Category info.
  - Logged actions.
  - User-groups.
- There's life beyond revision history!!

Pages-meta-history

- Most popular dump files in Wikipedia research works.
- Dump of 3 MediaWiki tables.
  - Page.
  - Revision.
  - Text.
- For every wiki page, all consecutive revisions are dumped.

Pages-meta-history

- General structure
- Example XML file in WikiDAT (furwiki).
  - Header.
  - Anonymous revision
  - Revision from registered user.
  - Other fields of interest.
    - Minor edit.
  - We can extract additional info from text content.

Pages-logging

- Dump of *logging* table in MediaWiki.
- Administrative and maintenance actions.
    - Example XML file in WikiDAT (simplewiki).
    - List of different actions Recorded.
- We can use namespace prefix in page title to annotate this info for every action.
- Sometimes, we can find actions specific to certain plug-ins.
    - "review" actions for flagged-revisions.

1.2 Data preparation and storage

- RSS to notify updates.
- Enwiki vs. rest of languages.
  - Huge size.
  - Multiple chunks (multiprocessing, clustering).
  - Hope for the best... get ready for the worst.
    - Missing revision users.
    - Missing (or empty) text.
    - Issues with charsets (e.g. got: in MySQL).

1.2 Data preparation and storage

- Extra metadata.
  - Revision parent id.
  - Revision length.
  - Information in text.
    - Tags (quality content, special templates).
      - Different languages.
      - See example in WikiDAT for FAs (later on).
    - Links (over time).
    - References.
    - Images, multimedia...

1.2 Data preparation and storage

- Tips and assessment.
  - Expected speed.
  - Configure your database.
  - Work in memory, if possible.
  - Don't underestimate the power of SSDs.
  - Multiprocessing better than multithreading.

1.2 Data preparation and storage

- Tips and assessment.
  - Hardware and operating system limitations.
    - Memory capacity.
    - Size of storage devices.
    - Multiprocessing in a single machine easier than clustering (map-reduce).
  - Working with dumps.

1.3 Available (FLOSS) tools

- Here be dragons

1.3 Available (FLOSS) tools

- Wikistats (Erik Zachte).
    - Perl scripts.
    - Overall metrics and trends for all Wikimedia projects.
    - Also provide some pre-computed data files (CSV format).
    - http://stats.wikimedia.org
    - WMF Labs reportcards.
        - http://reportcard.wmflabs.org/

1.3 Available (FLOSS) tools

- Pywikipediabot, python-wikitools, mwclient.
  - Interacting with MediaWiki API.
  - Reading and/or editing (user account).
  - http://www.mediawiki.org/wiki/Pywikipediabot
  - http://code.google.com/p/python-wikitools/

1.3 Available (FLOSS) tools

- Pymwdat (D. Chichkov, in Google Code).
  - Retrieve information from page dump files (SAX + threading).
  - Dumb diff algorithm to track differences between revisions (approx. vandalism detection).
  - Calculate some general metrics about pages, content and users.
  - http://code.google.com/p/pymwdat/

1.3 Available (FLOSS) tools

- StatMediaWiki and Wikievidens (Emijrp).
  - Creates graphics and scores to analyze the status and evolution of MediaWiki sites.
  - Wikievidens: comprehensive tool for dataset downloading, XML processing and analysis and visualization of general statistics.
  - http://statmediawiki.forja.rediris.es/index_en.html
  - http://code.google.com/p/wikievidens/

1.3 Available (FLOSS) tools

- WikiTrust (UCSC, parallel, cluster).
  - Focused on authorship and reputation.
  - Produces 3 types of metadata:
    - Revision where each word was introduced.
    - Author of each word.
    - To what extent the word was revised in subsequent edits (deletion or moves).
  - Equations to calculate author reputation based on authorship info.
  - Complex, requires clustering.

1.3 Available (FLOSS) tools

- Wikimedia utilities (A. Halfaker).
  - Example software to process dump files (in parallel, multiprocessing).
  - Can be extended to extract or calculate extra information or metadata.
  - To parallelize, we need the dump to be sliced in multiple chunks.
    - Currently, only enwiki.
  - https://bitbucket.org/halfak/wikimedia-utilities

1.3 Available (FLOSS) tools

- WikiDAT (Felipe Ortega, A. Halfaker).
  - Wikipedia Data Analysis Toolkit.
  - Integral solution, covers all phases of data analysis (retrieval, preparation EDA and example models).
  - Python, MySQL, R.
  - Support for our examples.
  - [[LINK TO GITHUB PROJECT]]

2. Conducting Wikipedia data analysis

2.1 Methodology

- Automate as many steps as possible.
- Interpretation of results, model evaluation and rebuilding cannot be automated.
- Steps.
  - Identify sources.
  - Retrieve and store information.
  - Preapre and clean data.
  - EDA.
  - Model building and interpretation.
  - Write your report or publish results.

2.1 Know your data

- The curious case of the timestamps.
  - Can we find two or more revisions for the same page with the same timestamp?
  - Can we find two or more revisions by the same user with the same timestamp?
- Importance of knowing our data and its generation process.
  - Improve data preparation.

2.1 Routinary tasks

- Keep data preparation in database.
  - In general, it is preferable to perform data preparation in the database.
  - Unless it renders impossible (for instance, in high-resolution analysis requiring clustering).
- Separate anonymous editors.
  - IP useless to track them accurately.
  - For example, the case in which Wikipedia accidentally banned edits from Quatar.

2.1 Routinary tasks

- Bots and extremely active editors.
  - Filter out edits from bots if you are intersted in human contributions.
  - But beware of extremely prolific wikipedians.
- Prepare for any missing fields.
  - Fill in the gaps wisely (imputation).
- Widespread definitions.
  - E.g. active and very active wikipedians.

2.2 FLOSS tools for data analysis

- Python.
    - NumPy, SciPy, matplotlib.
    - Scikit.learn.
- MySQL (or PostgreSQL).
- R programming language.
    - *De facto* standard for statistical computing.
    - +3,800 libraries with extended features.
    - http://r-project.org
- Refer to the companion guide for more info.

2.3 Case examples

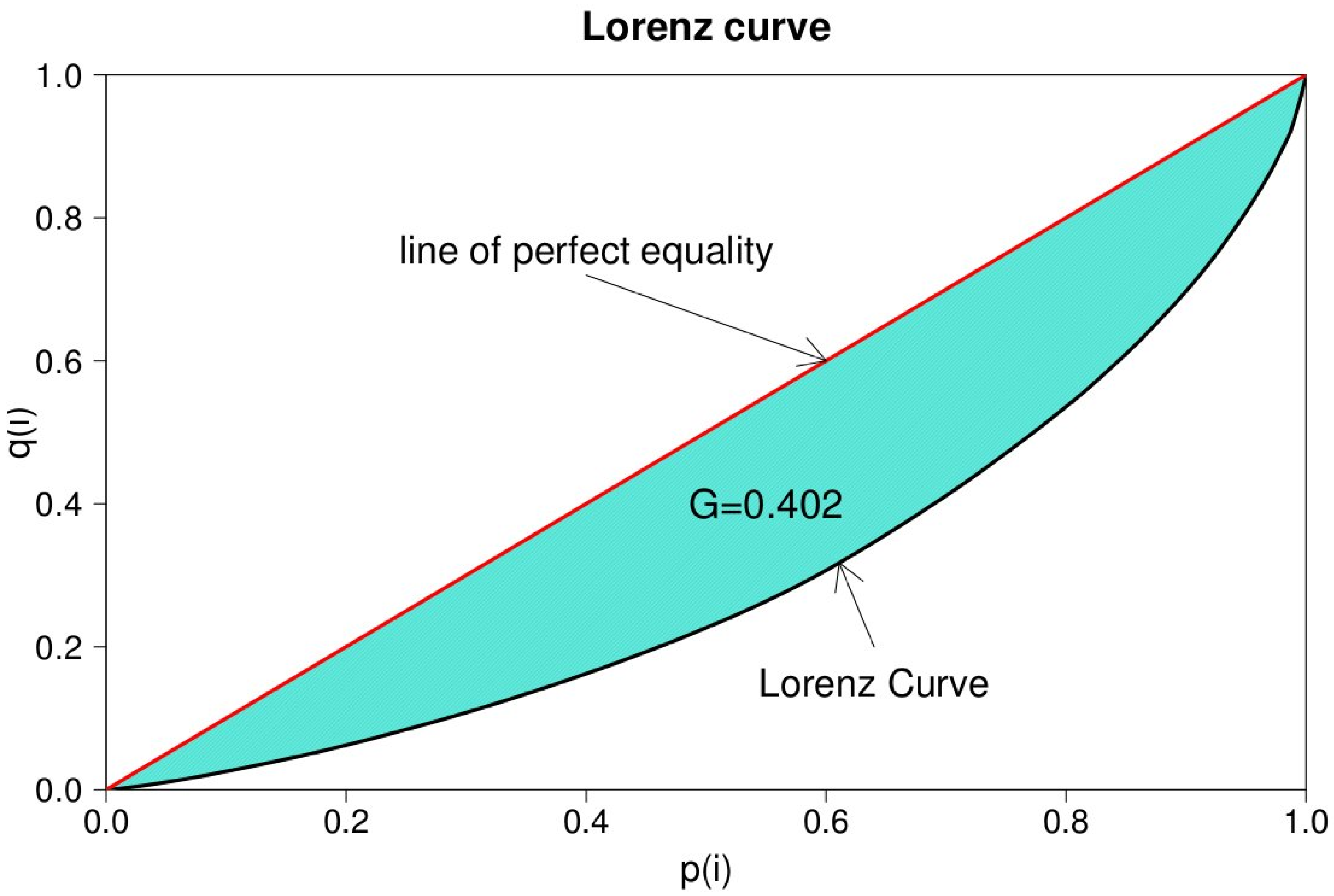
2.3.1 General statistics

- General statistics for a given month.
  - Intermediate tables created in DB.
  - CSV file produced in Python.
  - Loading data in R to complete analysis.
  - Example for August 2011.
- Involved R packages.
  - *Hmisc* and *car*.
- Directory *tools/activity* in WikiDAT.

2.3.2 The study of inequalities

- Analize inequality of contributions from registered users.
- Use Lorenz curve and Gini coefficient.
- R package *ineq*.
- > *install.packages(“ineq“, dep = T)*
- Load *revisions.RData* and *users.RData*.
- *Inequality* directory in WikiDAT.

2.3.2 The study of inequalities



2.3.3 Logged actions

- Case study: Simple English Wikipedia
  - Simplewiki
- Parse dump file *pages-logging*.
  - Prepared SQL file.
- Analize evolution of logged actions.
  - Folder *tools/logging* in WikiDAT.
  - User blocks and page protection.
- Seasonality and trend decomposition.

References

- WikiDAT repository on Github.
- Companion guide.
  - Sources on Github (Wikidat repository).
  - PDF version on Wikimedia commons (coming soon).
- R references.
  - R manuals and contributed documentation.
  - R in a Nutshell (O'Reilly, 2011).
  - Introductory statistics with R (Springer, 2008).