Web Content-Based Statistics: The Challenges Ahead

Fernando REIS

Web Intelligence Network Conference - From Web to Data
Gdansk, 4-5 February 2025

31/01/2025

European Commission

Challenges Overview

- Instability of the Web
- Duplication of objects
- Automatic information extraction
- Fakery and misinformation
- Representativeness

Instability of the Web

- Websites appear, disappear, or change
- Downtime and access restrictions
- Impact on continuity and time series consistency

- It’s unavoidable
- We need methods to address this instability
  - E.g. Chaining
    - Promissing, but we need to address breakdowns

Duplication of Objects

- A curse and a blessing
  - Duplicates lead to over-estimation of totals
  - Redundancy across websites, reduces impact of instability of the web
- Duplication happens across websites and within websites

- Possible solutions:
  - Restrict the web sources: eliminates the curse, but also the blessing
  - Increase the effectiveness of the deduplication
  - Surveys on web sources owners and statistical units (enterprises, individuals)

Automatic Information Extraction

- Need for automated methods (NLP, AI)
- Human annotation / labelling is **very** expensive
- Precision of latest AI developments (LLM) put algorithms at par with humans
- Trade-off between cost and precision of AI
- Measurement errors introduced by algorithms bias our statistics
- We must be able to measure the precision of the algorithms

- Solution(s):
  - We urgently need gold standards / test datasets to estimate precision using LLMs

Fakery and misinformation

- How fakery differs from noise – bias
- Intentional distortions targeting key variables
- Not much work done in official statistics

- Solutions:
  - Source validation and trustworthiness assessment
  - Detection using AI
  - Cross-validation with other data sources
  - Human expert oversight & hybrid approaches

Representativeness

- Coverage and selectivity
- Bias in web-based data (who is represented, who is not?)

- Solutions:
  - Estimation methods that correct selectivity – auxiliar information is required
    - Eurostat (2018) An overview of methods for treating selectivity in big data sources
  - We need specific solutions for specific use cases

Future Directions

- Continue developing methodologies
- Need for cross-disciplinary collaboration
- Investments in infrastructure and expertise

Q&A

- Questions? Remarks?