A Methodology and a Toolbox to Explore Dataset related to the Environmental Impact of HTTP Requests

Christophe Cérin, **Mathilde Jay**, Laurent Lefevre, Denis Trystram

LIPN, USPN
LIG, Grenoble
LIP Laboratory, Inria, ENS Lyon

BD4S'23 – 16/12/2023

Table of Contents

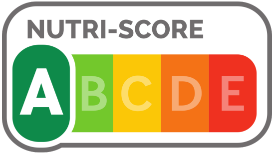
Part. 1 Introduction

- Yes, websites have a carbon footprint;
- From websitecarbon.com:
  - The internet consumes a lot of electricity: 416.2 TWh per year;
  - Globally, the average web page produces approximately 0.8 grams of CO2 equivalent per page view.
- From the International Energy Agency, in 2022:
  - Data Centers consumed between 240 and 340 TWh;
  - Data Transmissions accounted for 260 to 360 TWh;
  - Which represents 2 to 2.8% of the global energy demand in 2022 (ourworldindata.org)
- An historical tool to estimate the impact of visiting a website: EcoIndex.
- We question this tool and propose alternatives.

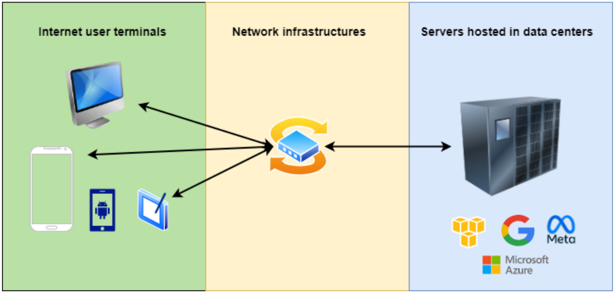
Introduction

Why and how does the index is calculated?

- Assume a representative database of HTTP requests (a collection, day after day of HTTP traces): HTTParchive (https://httparchive.org/);
- By measuring the environmental impact of the requests stored on it $\Rightarrow$ monitor the evolution of the web.



3-tiers technical model



Introduction

How does EcoIndex capture the 3-tiers model?

- Internet user terminals: the number of DOM (Document Object Model) elements; (DOM attribute)
- Network infrastructures: the weight of the page downloaded from the servers to the terminal (expressed in KB downloaded); (SIZE attribute)
- Servers in data centers: number of HTTP requests between terminal and servers. The number of HTTP requests hidden in the HTML page. (REQ attribute)
- Weights between attributes: DOM size weighs $\sim$ 3 times more than the page weight – the number of queries weighs $\sim$ 2 times more than the page weight. (Based on meta-studies)

$$EcoIndex = 100 - \frac{3 * F_{DOM}(DOM) + 2 * F_{REQ}(REQ) + 1 * F_{SIZE}(SIZE)}{6}$$

Part. 2 Criticisms of the EcoIndex – Statistics

Criticisms of the EcoIndex – Statistics

Criticisms

- We formulated three types of criticisms:
  1. Limitations inherent to the 3-tiers model: no life cycle analysis (LCA); but EcoIndex is "high level" and easy to understand by a non-expert;
  2. Limitations inherent to the calculation itself: a) do not consider variations over time for the weights b) the user's geographical location c) quantiles are hidden in the F functions: how they were determined because they determine the A-G scores?
  3. Limitations inherent in the attributes that make sense: a) the energy mix is not in the model b) 4/5G mobile or fiber?

- We formulated three types of criticisms:
  1. Limitations inherent to the 3-tiers model: no life cycle analysis (LCA); but EcoIndex is "high level" and easy to understand by a non-expert;
  2. Limitations inherent to the calculation itself: a) do not consider variations over time for the weights b) the user's geographical location c) quantiles are hidden in the F functions: how they were determined because they determine the A-G scores?
  3. Limitations inherent in the attributes that make sense: a) the energy mix is not in the model b) 4/5G mobile or fiber?
- We need, on the one hand, for the community to agree on these new attributes, and then we need to have associated computational methods able to process a large number of attributes.

Criticisms of the EcoIndex – Statistics
Statistics

- We designed a Jupyter Notebook (https://github.com/christophe-cerin/Ecoindex-Revisited/blob/main/analysis_mj.ipynb);
- To better understand the April 2022 extraction of the HTTParchive (108k URLs); Examples: distribution of DOM, REQ, SIZE, and EcoIndex score;
- Check/validate/invalidate our assumptions.

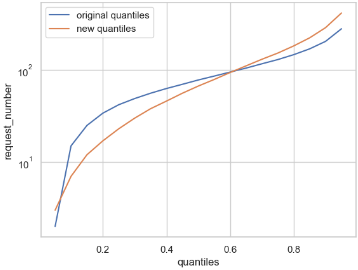
Criticisms of the EcoIndex – Statistics

Statistics

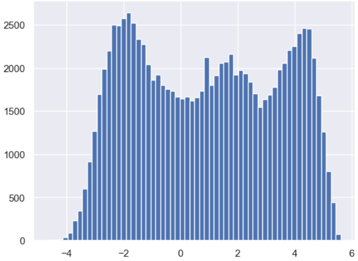
- We designed a Jupyter Notebook (https://github.com/christophe-cerin/Ecoindex-Revisited/blob/main/analysis_mj.ipynb);
- To better understand the April 2022 extraction of the HTTParchive (108k URLs); Examples: distribution of DOM, REQ, SIZE, and EcoIndex score;
- Check/validate/invalidate our assumptions.
- Examples of findings:
  - The EcoIndex score has a uniform distribution on our dataset, confirming that our dataset is diverse. The numbers of values at the extreme part of the spectrum are few, which is expected and confirms that the definition of the Ecoindex is relevant.
  - The quantiles from our dataset extraction are similar to the original quantiles $\Rightarrow$ stability;

Criticisms of the EcoIndex – Statistics

Examples of statistics (over 108k of HTTP requests)



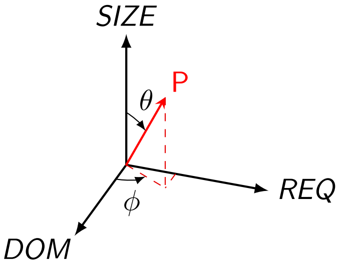
(a) Quantile's similitude for REQ parameter – y-axis is the number of HTTP/HTTPS in the returned HTML page.



(b) With 60 bins for the histogram, we observe that the difference in values for the two EcoIndex is at a maximum of 2500 – Difference positive or negative but low – y-axis is the number of EcoIndex scores with that difference.

Part. 3 Proposals for new definitions

- Three objectives: a) still compute a score b) add (many) more attributes c) no weighting;
- Example with the 3-D space: projection of the point P on the line + the (normalized) position is the score.
- Key data science technique: dimensionality reduction.



- Many techniques exist: Random projection, Locality Sensitive Hashing (LSH), computing the collinearity (we keep the k most collinear points with the given point, and we compute a sum over an average for the dimensions → idea of similarity);
- We experimented with all of them;

Proposals for new definitions

Performance evaluation (Examples)

Table: RANDOM PROJECTION VERSUS COLLINEARITY

Average Root Mean Square Error:	27.01
Min Root Mean Square Error:	0.009
Max Root Mean Square Error:	79.45

Table: RANDOM PROJECTION VERSUS LSH-KNN:

Average Root Mean Square Error:	26.94
Min Root Mean Square Error:	0.019
Max Root Mean Square Error:	71.84

- The new approaches seem significantly different since the RMSEs are sparse;
- Which one to choose? The energy criterion.

Proposals for new definitions

Energy consumption with CodeCarbon on DELL Elitebook (core i7, 32GB RAM)

Table: Energy and CO2 emission for the exploration of 100k URLs

	EcoIndex	Random proj.	LSH	Collinearity
Energy	0.000121 kWh	0.000011 kWh	X	0.036015 kWh
CO2	5.806e-05 kg	2.676e-08 kg	X	0.0172348 kg

Notice the historical EcoIndex is not sensitive to the location.

Table: Sensitivity analysis regarding the location

Location	URL	GRADE	CO2
JAPAN	`https://www.nii.ac.jp/en/faculty/digital_content/andres_frederic/`	C	3.317e-05 kg
JAPAN	`https://research.nii.ac.jp/~andres/official/content_e.html`	A	3.154e-06 kg
FRANCE	`https://www.nii.ac.jp/en/faculty/digital_content/andres_frederic/`	C	1.220e-05 kg
FRANCE	`https://research.nii.ac.jp/~andres/official/content_e.html`	A	1.280e-06 kg

Proposals that DO NOT reduce to a score

Self Organizing Map (SOM)

- SOM: an unsupervised machine learning technique used to produce a low-dimensional (typically two-dimensional) representation of a higher-dimensional data set while preserving the topological structure of the data.
- Objectives: a) many attributes; b) a point in the high dimensional space maps to a point in the 2-D space, and, visually, we get similar points (clustering); c) Consider the location; d) No weighting;
- Idea: facilitate or guide intra/extra-clusters interpretations;

Proposals that DO NOT reduce to a score

Self Organizing Map (SOM)

- SOM: an unsupervised machine learning technique used to produce a low-dimensional (typically two-dimensional) representation of a higher-dimensional data set while preserving the topological structure of the data.
- Objectives: a) many attributes; b) a point in the high dimensional space maps to a point in the 2-D space, and, visually, we get similar points (clustering); c) Consider the location; d) No weighting;
- Idea: facilitate or guide intra/extra-clusters interpretations;
- We combine datasets (ENEDIS, and ARCEP) to build a 14 attributes model (with categorical data):
  - URL, DOM, REQ, SIZE, EcoIndex, city name, and energy mix of the city, page loaded in less than 10s. . .

Proposals that DO NOT reduce to a score

Self Organizing Map (SOM) with 14 attributes

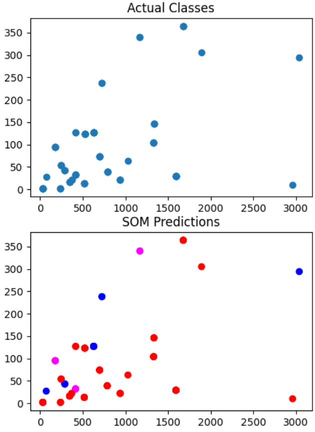


Figure: Topological map with public data from Enedis, Arcom

Part. 4 Conclusion and Future work

Conclusion and Future work

Summary

- This paper discussed the EcoIndex, an environmental score for an HTTP request, and some of its limitations;
- We also compared qualitatively alternative approaches anchored in Data Sciences;
- Most promising approaches: a) Random projection (better, regarding the CO2 emission) b) SOM (better, regarding the interpretability: understanding the inner workings of the models);
- Toolbox available on https://github.com/christophe-cerin/Ecoindex-Revisited;
- We call on the entire IT community and the citizens to mobilize on the issues raised in this article. Crossing, in particular, competencies in Architecture, Systems, Networks, and Learning is necessary.

Conclusion and Future work

Future works – Adjacent work in the HPC context

- Idea of use: recommendation system for the citizens to decrease the emission for IT;
- Move from the exploration phase to the modeling phase, and then use the model(s) in practice; (Data Science projects go like this!)
- In parallel, we conduct works on more fine-grained approaches to model/estimate the electricity consumed and the environmental impact of ICT
  - [Jay et al., 2023]: Comparing Software-based Power Meters
  - [Berthelot et al., 2023]: LCA-based methodology to estimate the environmental impact of Generative AI services, illustrated on Stable Diffusion
  - Evaluating AI training scenarios depending on where they are executed (edge, cloud, in-between)

Thank you for your attention. Do you have any questions?

Bibliography

Berthelot, A., Jay, M., Lefevre, L., and Caron, E. (2023).
Estimating the environmental impact of Generative-AI services using an LCA-based methodology.
working paper or preprint.

Jay, M., Ostapenco, V., Lefèvre, L., Trystram, D., Orgerie, A.-C., and Fichel, B. (2023).
An experimental comparison of software-based power meters: focus on CPU and GPU.
In *CCGrid 2023 - 23rd IEEE/ACM international symposium on cluster, cloud and internet computing*, pages 1–13, Bangalore, India. IEEE.