Internal mesh optimization Semantic linking and siloing Big data

DUPREY Stéfan

Cdiscount

Plan

1. Introduction

Objective : give you an overview of the meshing optimization done at Cdiscount, the first french e-commerce web site

- Heuristic based optimization algorithm to push specific products
- An e-commerce pitch
- Semantic similarity and constraining for shrinking our universe
- Big data implementation

2. Some notations

Let $N \in \mathbb{N}$ be the number of nodes in our mesh.
Let $(X_i)_{i \in \{1,\dots,N\}}$ be the vertices (URLs) of our oriented graph.
Let $(G_{ij}) \in \{0,1\}^{N \times N}$ be the adjacency matrix of our oriented graph.
Let here define f, a given data per URL, which gives a potentiality metrics for our vertices.

$$\begin{array}{rccl} f & : & (X_i)_{i \in \{1,\dots,N\}} & \rightarrow & \mathbb{R}^+ \\ & & x & \mapsto & f(x) \end{array} \tag{1}$$

3. In-rank computation

We restrain the universe to our site where we compute the standard page-rank.

Initialization :

$$\forall u \; PR\,(u) = \frac{1}{N} \tag{2}$$

Iterative computation :

$$PR\,(u) = \frac{(1-c)}{N} + c \times \sum_{v \to u} \frac{PR\,(v)}{card\,(\{v \to u\})} \tag{3}$$

4. In-rank computation

We restrain the universe to our site where we compute the standard page-rank.

Initialization :

$$\forall u \, PR(u) = \frac{1}{N} \tag{4}$$

Iterative computation :

$$PR(u) = \frac{(1-c)}{N} + c \times \sum_{v \to u} \frac{PR(v)}{card(\{v \to u\})} \tag{5}$$

5. Exhaustive brute force doesn't work

For a $N = 10^6$ millions URLs web site, we have 2^{N^2} with a 2048 bits mantissa, 256 bits exponent $2^{10^{6^2}} =$

9.5762442314927432848050594956989483747127095675192905698213

128517073583274396016675898714705184143146468453752442806484

690561169975498415015777492655947375270159476651418975300707

658547568802353384879419803574730952480197774380552040662758

127609571333683703207910070247048194459504686986124786492353

387550318495241621572271925127288273993787778380450774809611

395810191417363401889038757182279484019203870177413318113073

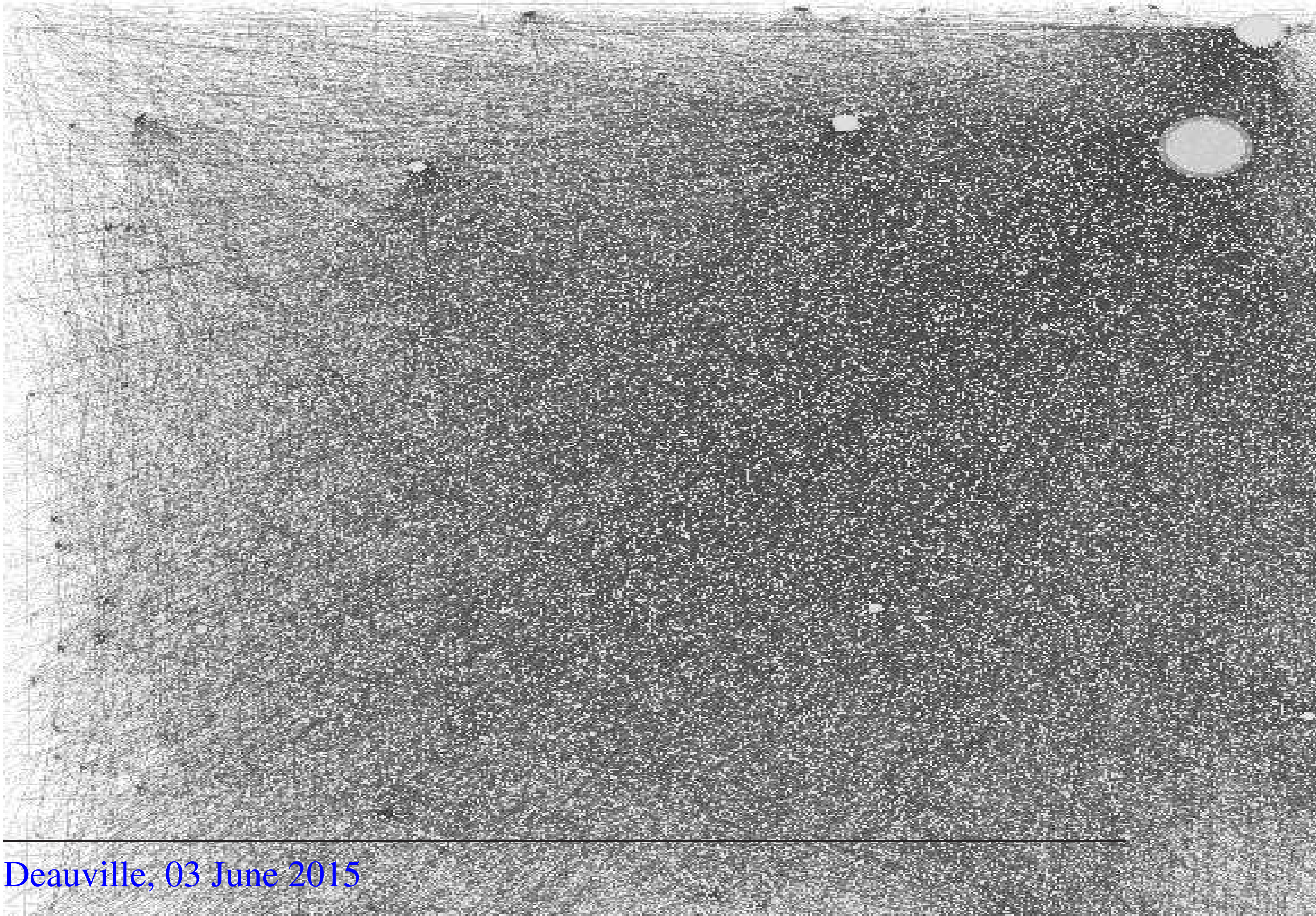
911418463615759647977538478560166958988721048687854280187283

661925937530017243461145905573802314471888491758757162677684

017424597014433418179115289463552630751896559312213624470617

4533250565836008e+301029995663

6. Picturing our smallest store : jewelery



7. Heuristic based algorithm

We want here to optimize the adequation of our mesh (X_i) to our potentiality vector f.

$$\max_{(G_{ij})\in\{0,1\}^{N\times N}} \left\{ \sum_{i=1}^{N} trafic\,(X_i) \times pageRank(X_i) \right\} \tag{6}$$

8. Genetic algorithm

What a genetic algorithm

- Genetic algorithm mimics evolutionary biology to find approximate solutions to optimization problems
- Start with an initial generation of candidate solutions that are tested against the objecive function (fitness of the individual)
- Subsequent generations evolve from the first through selection, crossover and mutation
- The individual that best minimizes the given objective is returned as the ideal solution

Why a genetic algorithm

- Lots of local minima to avoid
- Non continuous universe, constraints and objective
- Problem with noise and non-smooth data