Internal mesh optimization Semantic linking and siloing Big data

DUPREY Stéfan

Cdiscount

Plan

1. Introduction
2. Some notations
3. In-rank computation
4. In-rank computation
5. Exhaustive brute force doesn’t work
6. Picturing our smallest store : jewelery
7. Heuristic based algorithm
8. Genetic algorithm

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.

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

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 :

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

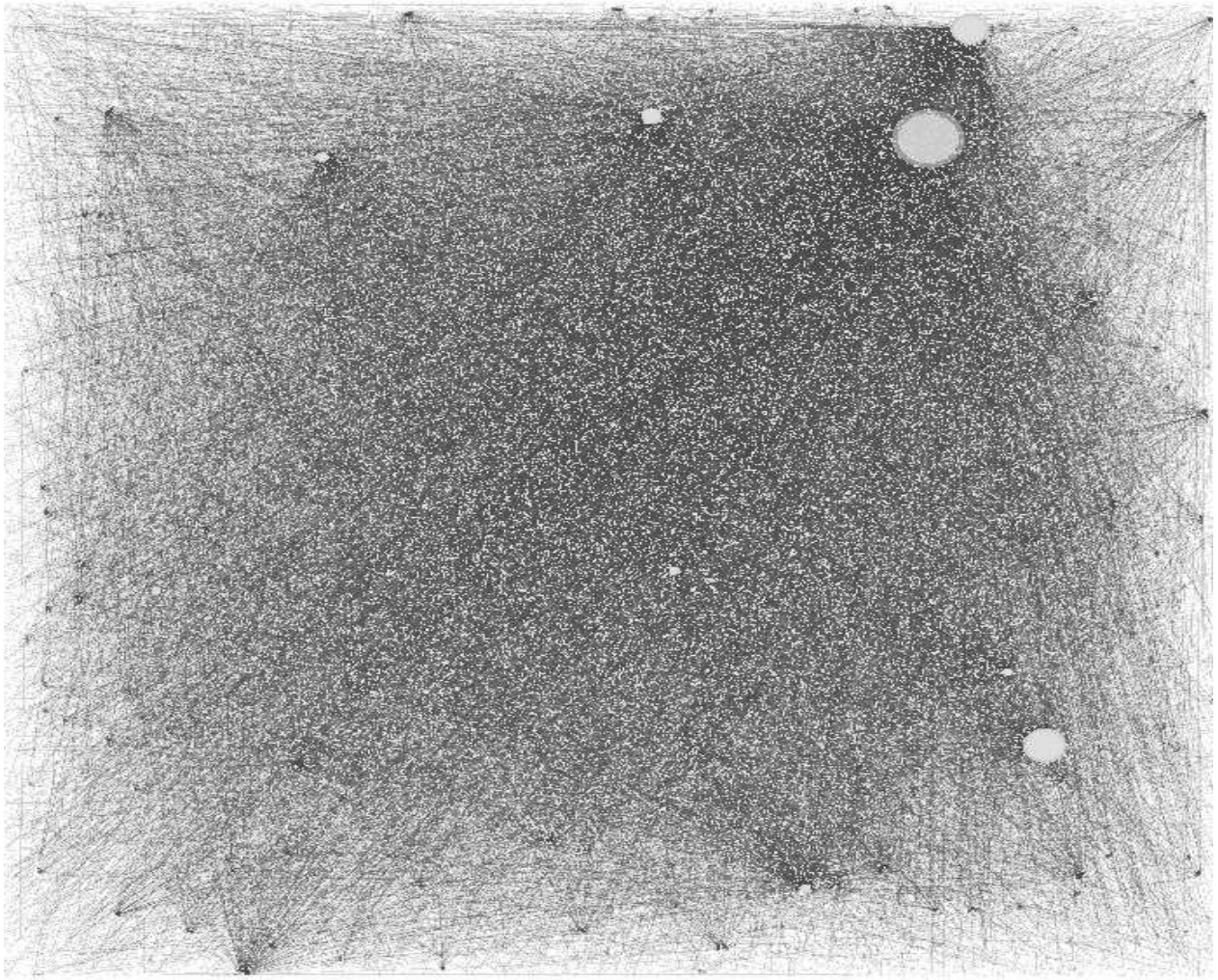
5. Exhaustive brute force doesn't work

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

$2^{10^{62}} =$

9.576244231492743284805059495698948374712709567519290569821312851707358327439601667589871470518414314646845375244280648469056116997549841501577749265594737527015947665141897530070765854756880235338487941980357473095248019777438055204066275812760957133368370320791007024704819445950468698612478649235338755031849524162157227192512728827399378777838045077480961139581019141736340188903875718227948401920387017741331811307391141846361575964797753847856016695898872104868785428018728366192593753001724346114590557380231447188849175875716267768401742459701443341817911528946355263075189655931221362447061745332505658360008e+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