Internal mesh optimization
Semantic linking and siloing
Big data

DUPREY Stéfan

Cdiscount

Plan

1. Introduction
2. Heuristic
3. Metaheuristics
4. Shrinking our universe with siloing and semantic similarity
5. Same problem with an e-commerce flavor
6. Prototyping example
7. Big data implementation : Neo4j and Spark GraphX

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. Heuristic

Some notations

Let $N \in \mathbb{N}$ be the number of nodes in our mesh.

Let $(X_i)_{i\in\{1,\ldots,N\}}$ 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.

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

In-rank

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 :

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

Our heuristic

We want here to optimize the adequacy of our mesh (X_i) to our potentiality vector f. We here postulate the following heuristic to assess the relevance of a mesh :

$$\max_{(G_{ij})\in\{0,1\}^{N\times N}} \left\{ \sum_{i=1}^{N} trafic\,(X_i) \times pageRank(X_i) \right\} \tag{4}$$

3. Metaheuristics

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

$2^{10^{6^2}} =$

9.5762442314927432848050594956989483747127095675192905698213128517073583274396016675898

7147051841431464684537524428064846905611699754984150157774926559473752701594766514 18975

3007076585475688023533848794198035747309524801977743805520406627581276095713336837 03207

9100702470481944595046869861247864923533875503184952416215722719251272882739937877 78380

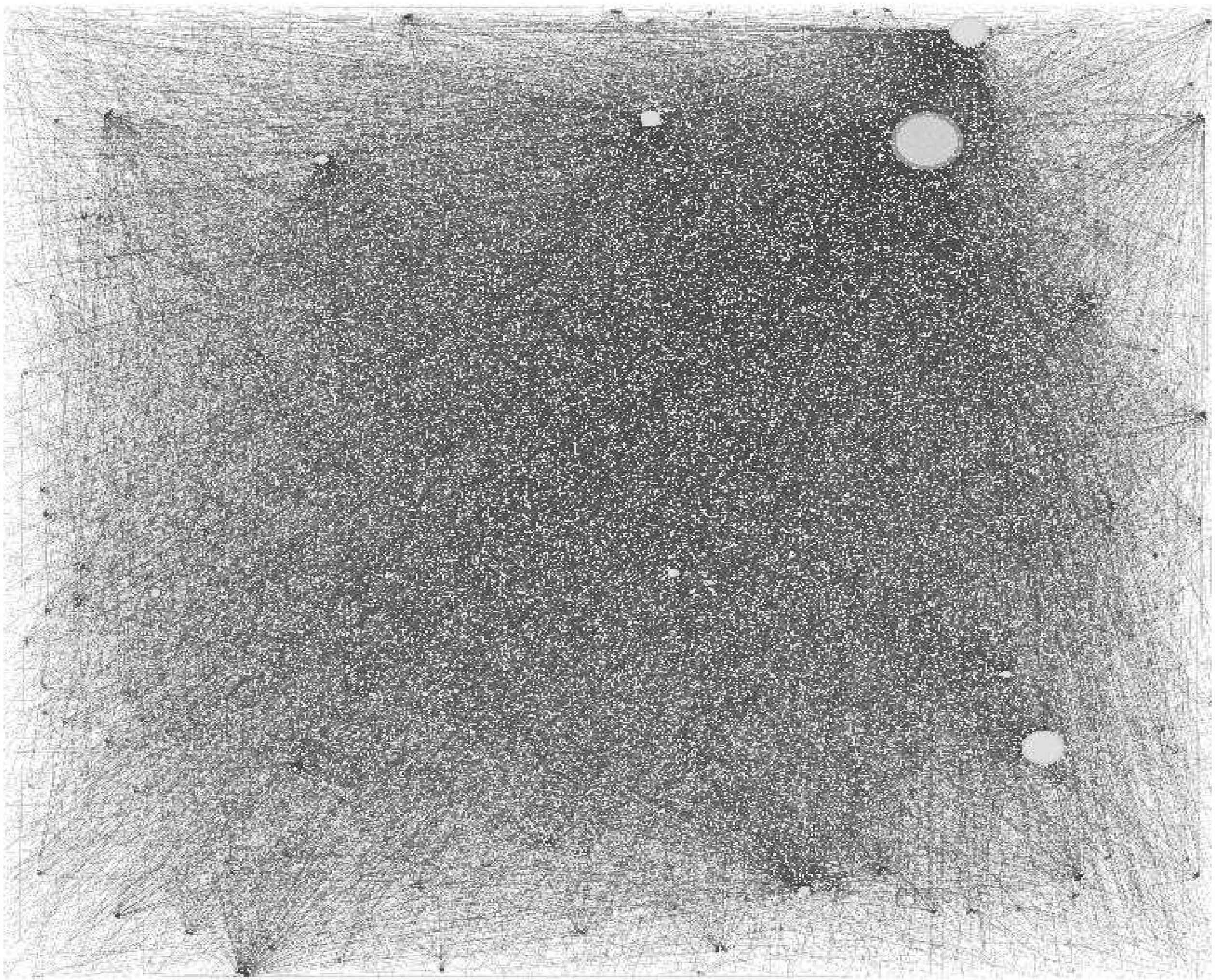
4507748096113958101914173634018890387571822794840192038701774133181130739114184636 15759

6479775384785601669589887210486878542801872836619259375300172434611459055738023144 71888

4917587571626776840174245970144334181791152894635526307518965593122136244706174533 25056

5836008e+301029995663

Picturing our smallest store : jewelery



Metaheuristics

We here have to maximize over all possible graphs. We have to find a clever global optimization methodology : metaheuristics such as global search, multistart, particle swarm, simulated annealing or genetic algorithm are all good candidates.

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 objective 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

Cleverly evolving through our universe

Child spawning from 2 parents crossover

(0 0 0 0 1 0 1 1 1 1 0 1 0)

(1 1 0 1 1 0 1 1 1 1 1 0 0)

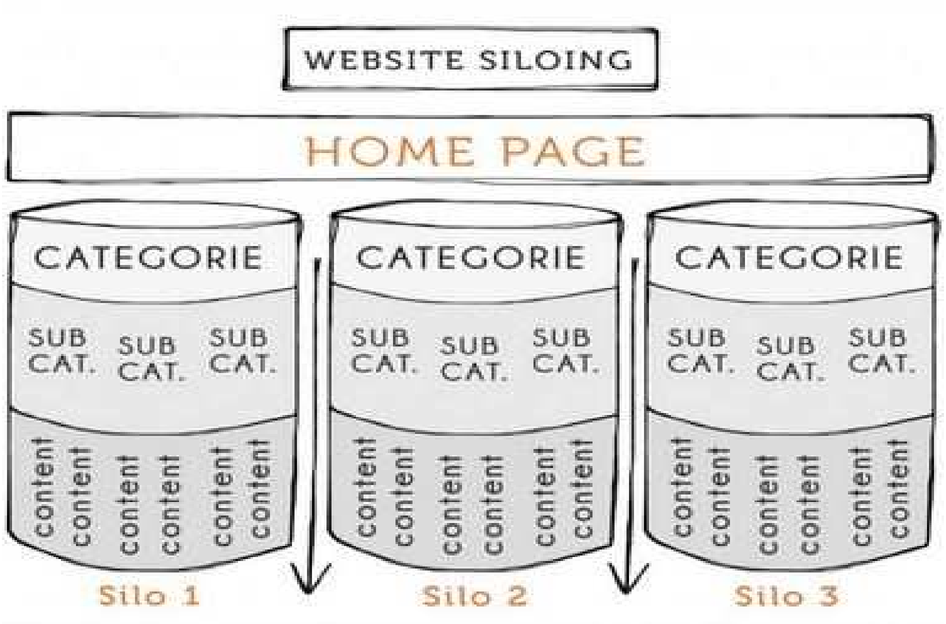
(* * 0 * 1 0 1 1 1 1 * * 0)

Mutation of an individual

(1 1 0 * 1 0 1 1 * 1 1 0 0)

4. Shrinking our universe with siloing and semantic similarity

Siloing and links categorizing



metalist1/metalist2 metalist1/list1

(001010 .. 00110 ..... 10100)

Semantic similarity

We slack our universe by allowing links only between semantically similar URLs :

$$\max_{(G_{ij})\in\{0,1\}^{N\times N}\ G_{ij}=0\ if\ CS(ij)\leq t} \left\{ \sum_{i=1}^{N} f\left(X_i\right) \times PR(X_i) \right\} \tag{5}$$

where $CS(ij)$ is a semantic distance between the two linked pages i and j. $CS(ij)$ can be defined very easily as the scalar product of the tf/idf vectors of the product descriptions whose weight are defined by the well known formula :

$$w_{ik} = \frac{tf_{ik} \log\left(\frac{N}{n_k}\right)}{\sqrt{\sum_{k=1}^{t}(tf_{ik})^2 \log\left(\frac{N}{n_k}\right)^2}} \tag{6}$$

5. Same problem with an e-commerce flavor

$$\sum_{i \in Keywords} SV\,(i) \times CTR\,(position\,(i)) \times CR\,(i) \times P\,(i) \qquad (7)$$

where $position\,(i)$ is the estimated position in search engine results coming from the modification of our new mesh,

$SV\,(i)$ is the search volume for the keywords i estimated by the search engine.

and $CTR\,(i)$ is the click through rate for an URL at the position place in the search engine results.

6. Prototyping example

$$(X_i) = \{'home','metalist1','metalist2',$$

$$'list1','list2','list3','list4',$$

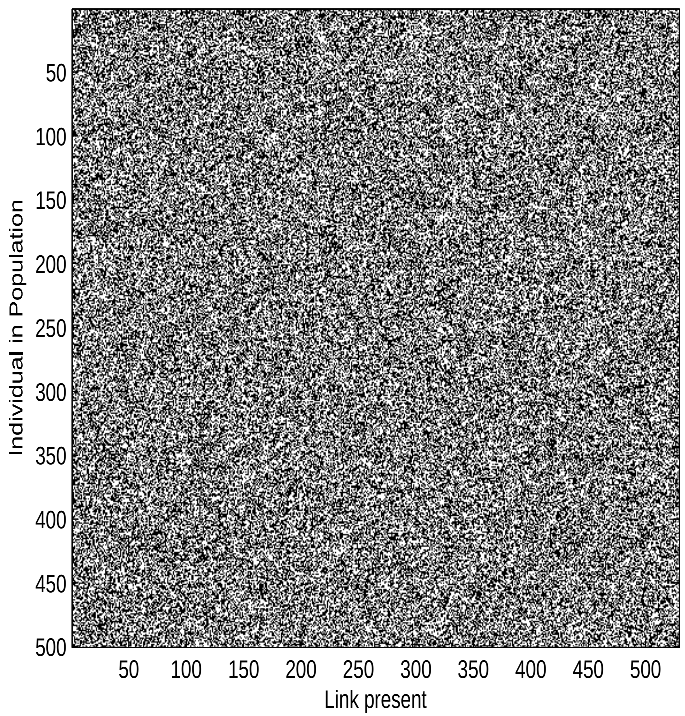
$$'product1','product2',...,'product32'\};$$

$$f\,(X_i) = [100; 80; 55;$$

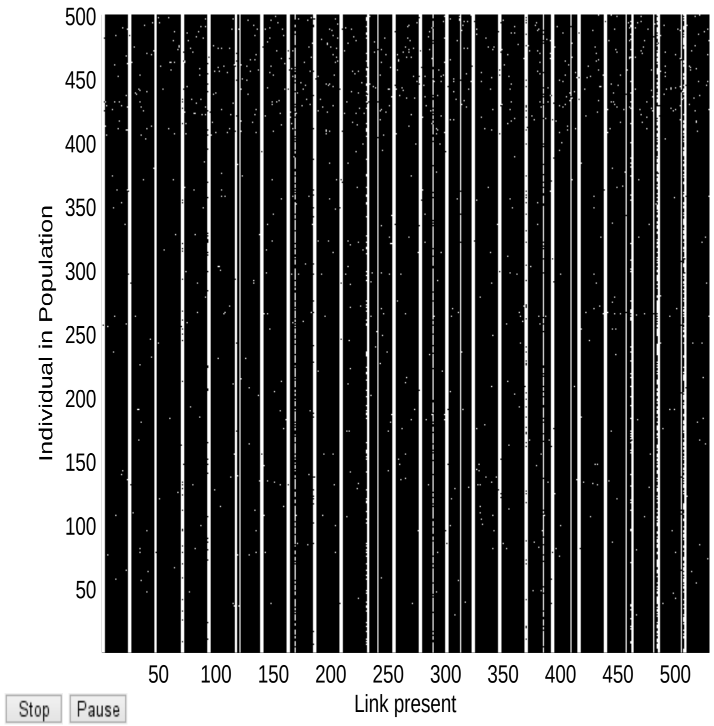
$$40; 35; 25; 20;$$

$$4; 5; ...; 3]$$

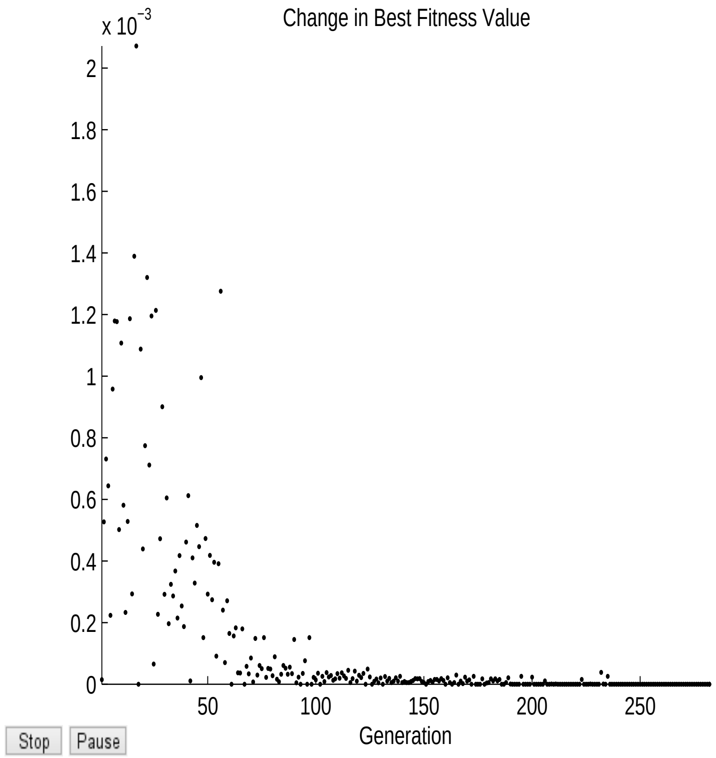
Initial population



Population at convergence

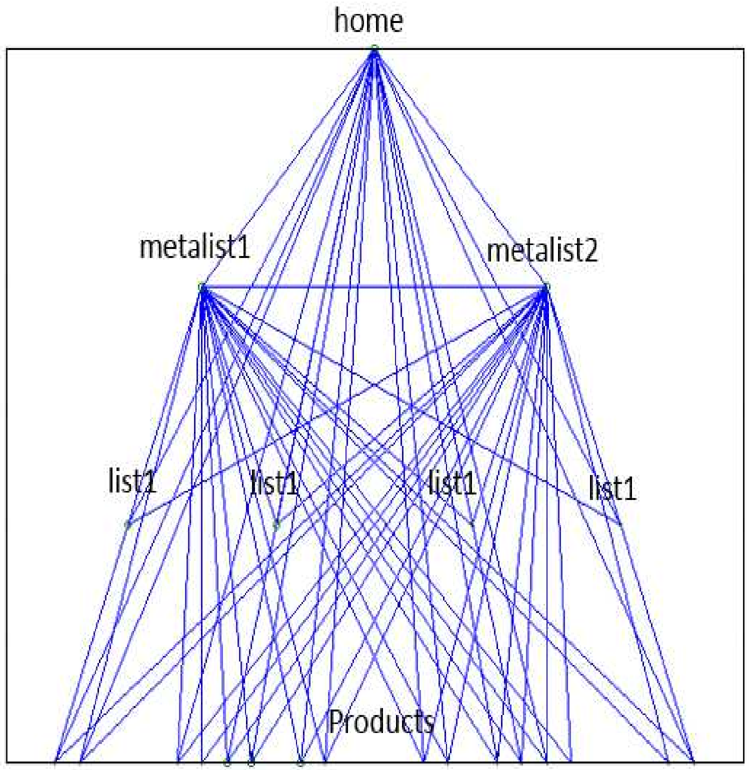


Best chromosome fitness function evolution



The structure found is as we expected arborescent : home linking to metalists, metalists linking to lists and lists to products.

Optimal site structure


home
metalist1
metalist2
list1
list1
list1
list1
Products

7. Big data implementation : Neo4j and Spark GraphX