Relembrando...

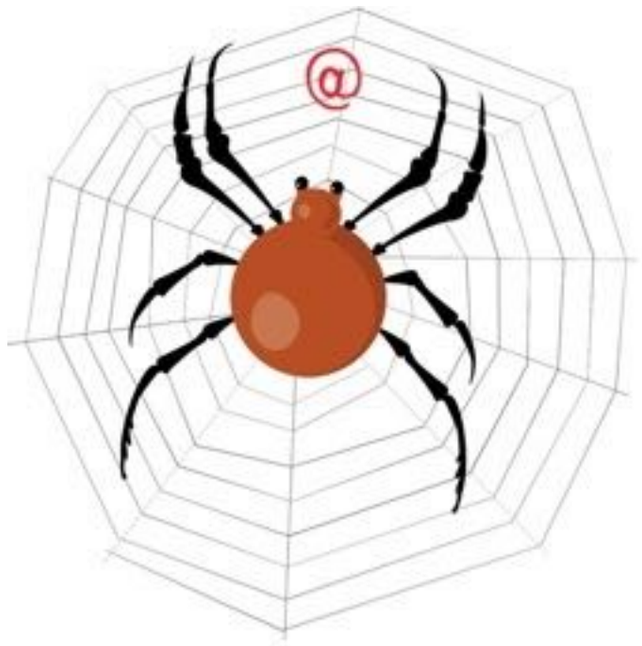
Domínio

Smartphones

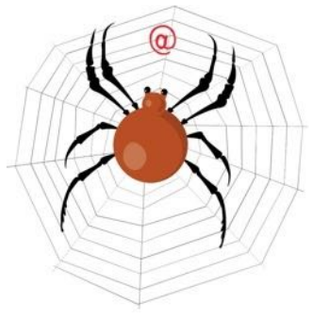
Sites

- Americanas
- Extra
- Submarino
- Nagem
- Saraiva
- Casas Bahia
- Ricardo Eletro
- Ponto Frio
- Cissa Magazine
- ShopTime

Crawler



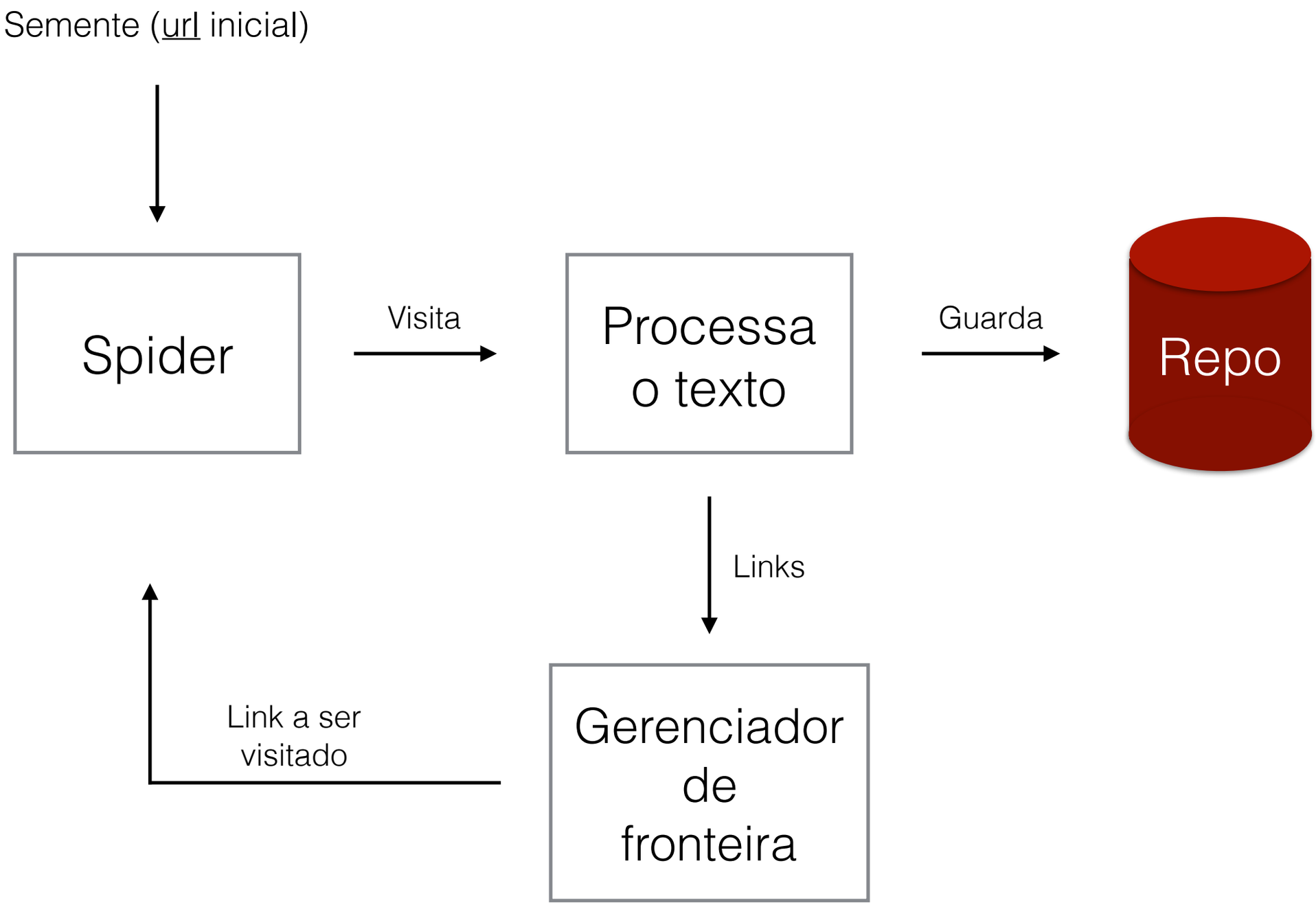
Roteiro



- Crawlers implementados
- Comparativo e conclusões

Crawlers implementados

Crawler


Semente (url inicial)
Spider
Visita
Processa o texto
Guarda
Repo
Links
Gerenciador de fronteira
Link a ser visitado

Crawler Busca em Largura

Gerenciador
de
fronteira

- Adiciona novos links no final da lista

Crawler Heurístico Sem Peso

Gerenciador
de
fronteira

- Adiciona novos links que contém alguma palavra relevante na âncora ou na URL no início da lista

- Os que não tiverem, vão pra o final

Crawler Heurístico Com Peso

Gerenciador
de
fronteira

- Calcula um peso pra o link baseado na presença de palavras heurísticas na âncora ou na URL

- A lista é ordenada e o próximo link visitado é o que tem maior peso

Crawler Heurístico Com Peso

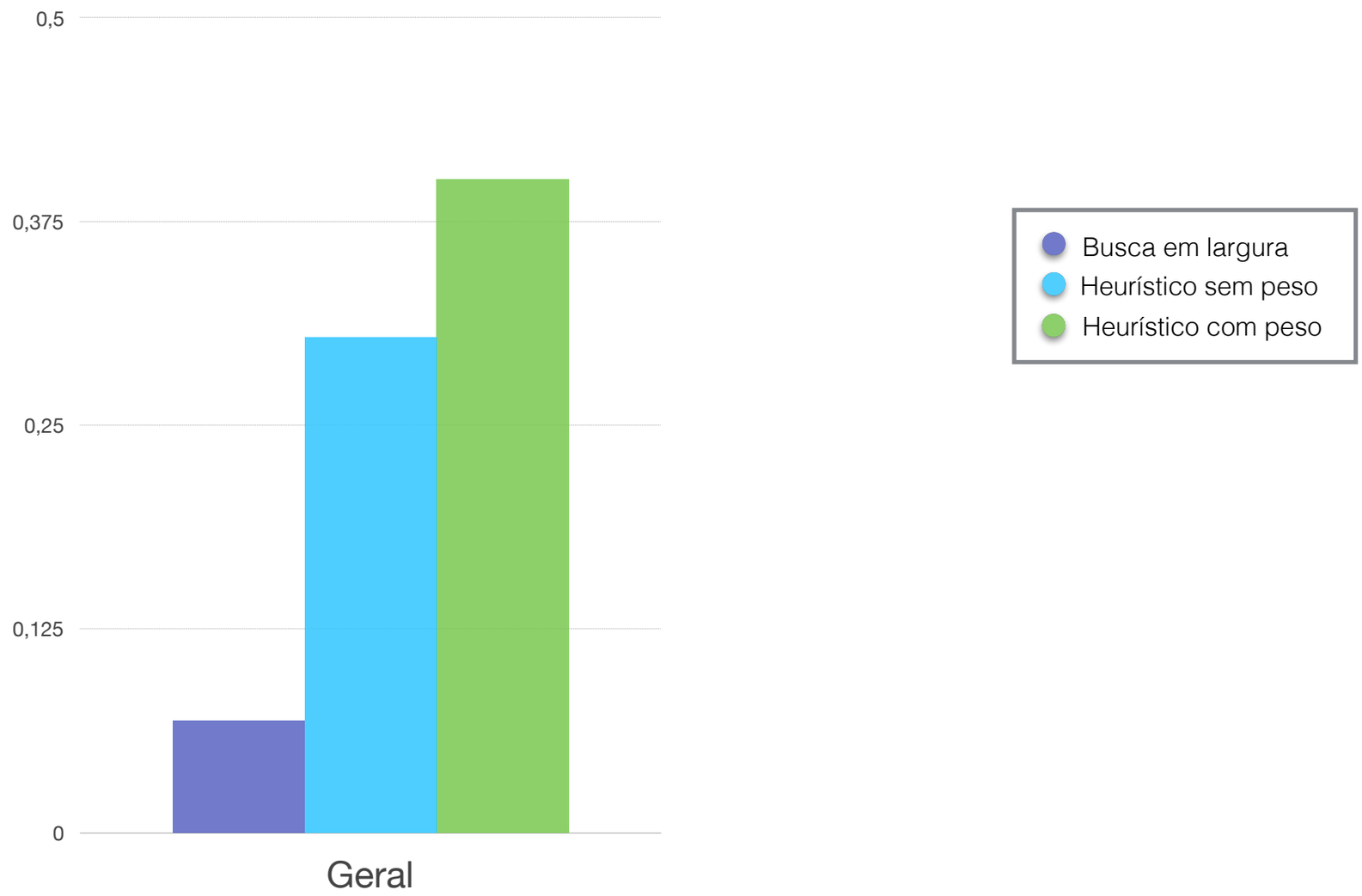
- Conjunto de palavras e pesos utilizado

```
samsung 4
apple 4
motorola 4
positivo 4
produto 4
desbloqueado 5
smartphone 2
celular 2
review -2
categoria -3
derivacao -3
order -3
filtro -5
avaliacao -7
page -7
capa -8
acessorio -8
linha -10
```

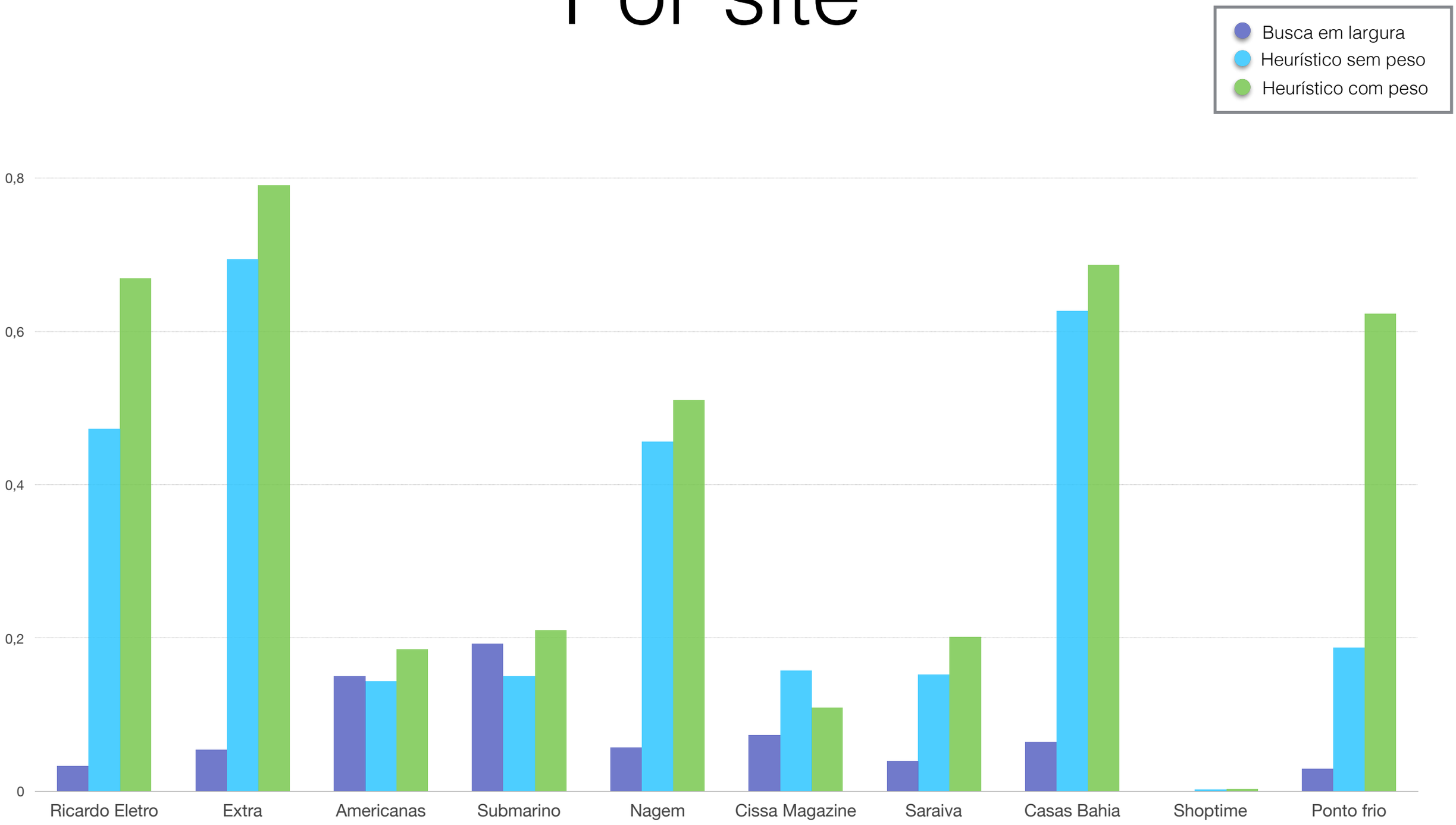
Comparativo e conclusões

(Harvest ratio)

Geral


0,5
0,375
0,25
0,125
0
Geral
Busca em largura
Heurístico sem peso
Heurístico com peso

Por site


Busca em largura
Heurístico sem peso
Heurístico com peso
0,8
0,6
0,4
0,2
0
Ricardo Eletro
Extra
Americanas
Submarino
Nagem
Cissa Magazine
Saraiva
Casas Bahia
Shoptime
Ponto frio

Considerações

- O site shoptime teve um resultado muito ruim por usar URL's relativas em seus htmls. Como o crawler sempre checa se a URL encontrada está no mesmo domínio da semente, as URL's das páginas dos smartphones nunca eram visitadas.

- Alguns sites tiveram uma densidade de produtos muito maiores porque possuem muito mais aparelhos celulares e smartphones em relação a outros, que só tinham modelos das principais marcas (e.g. Motorola, Apple e Samsung)