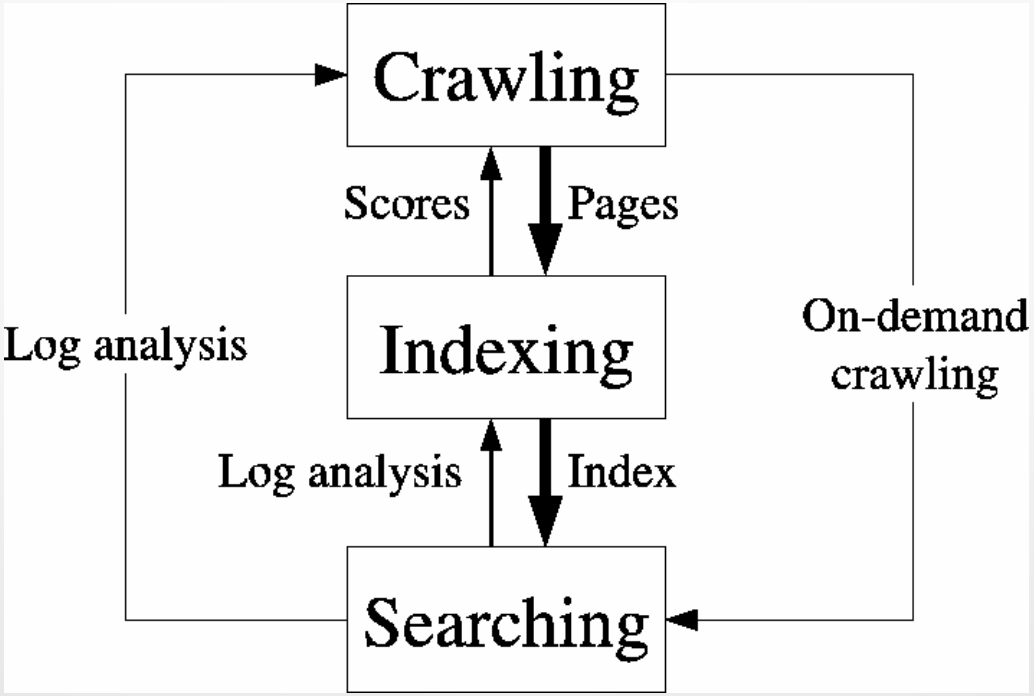
Uma Heurística para Web Crawlers

Alberto Hideki Ueda
ueda@dcc.ufmg.br

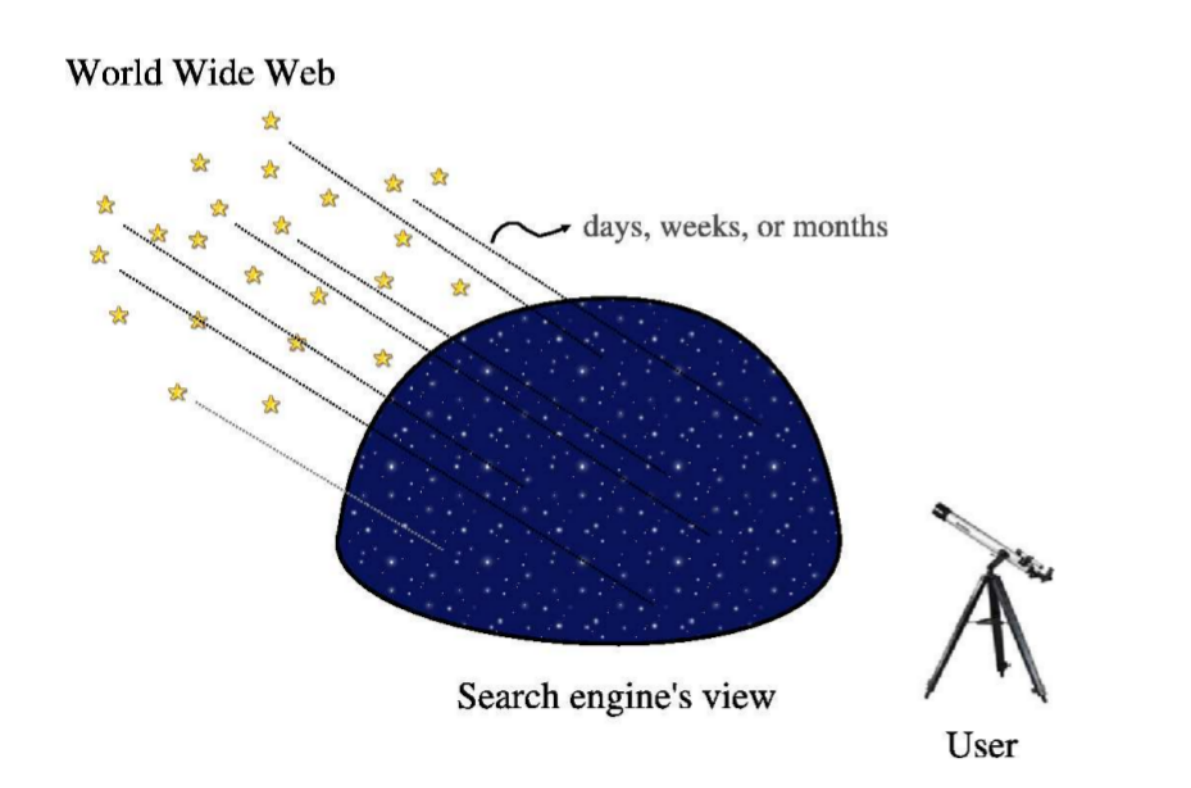
O Problema

- Primeiro passo para a implementação de máquinas de busca modernas
- Estima-se que hoje mais de 10% das visitas a Websites sejam feitas por coletores [Nielsen04]



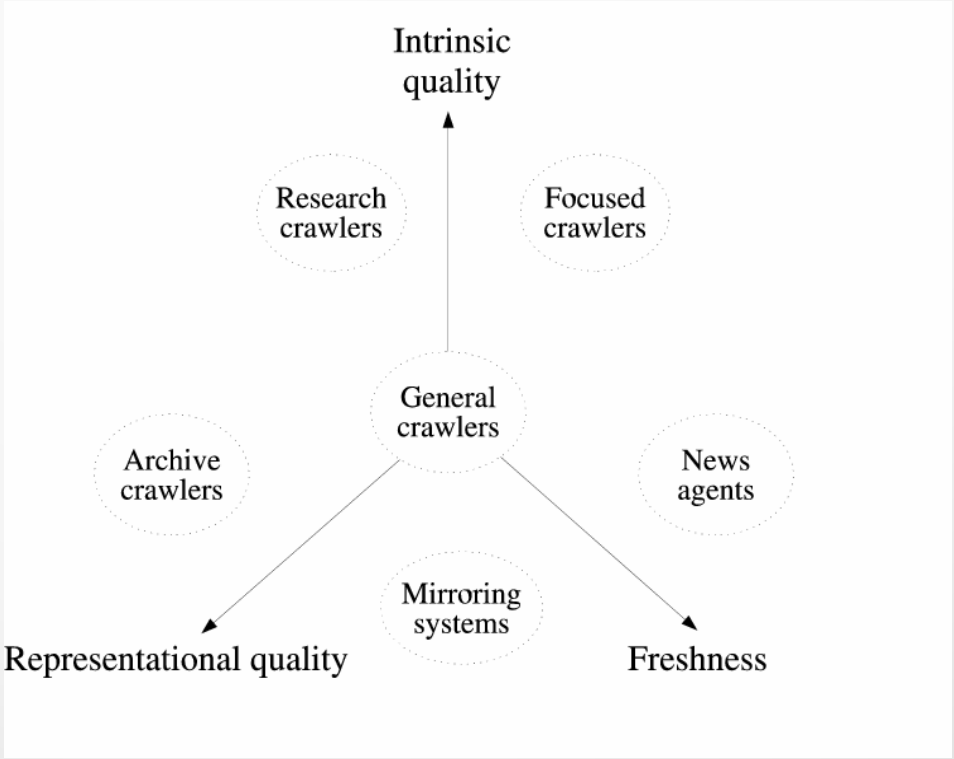
O Problema

- Como coletar a *Web*?
- Critérios: **qualidade**, **tempo**, memória

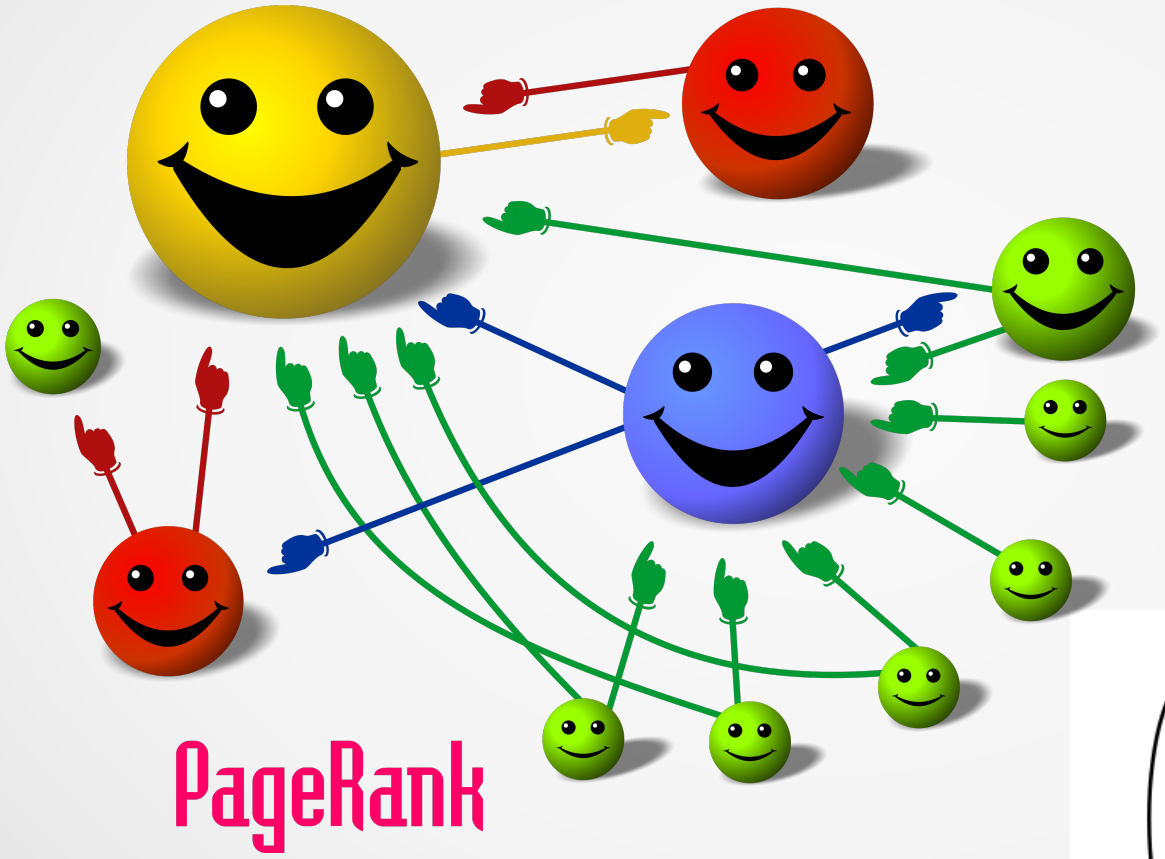


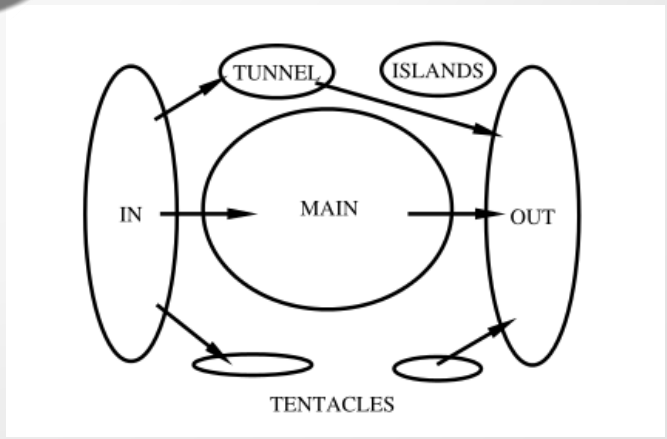
O Problema

- Nível de cobertura das principais máquinas de buscas existentes: entre 58% e 76% da *Web* [Gulli2005]
- Custo estimado para coletores de larga escala: US$ 1,5 M [Craswell2004]
- **Inviabilidade da Solução Exata**



Modelagem em Grafo





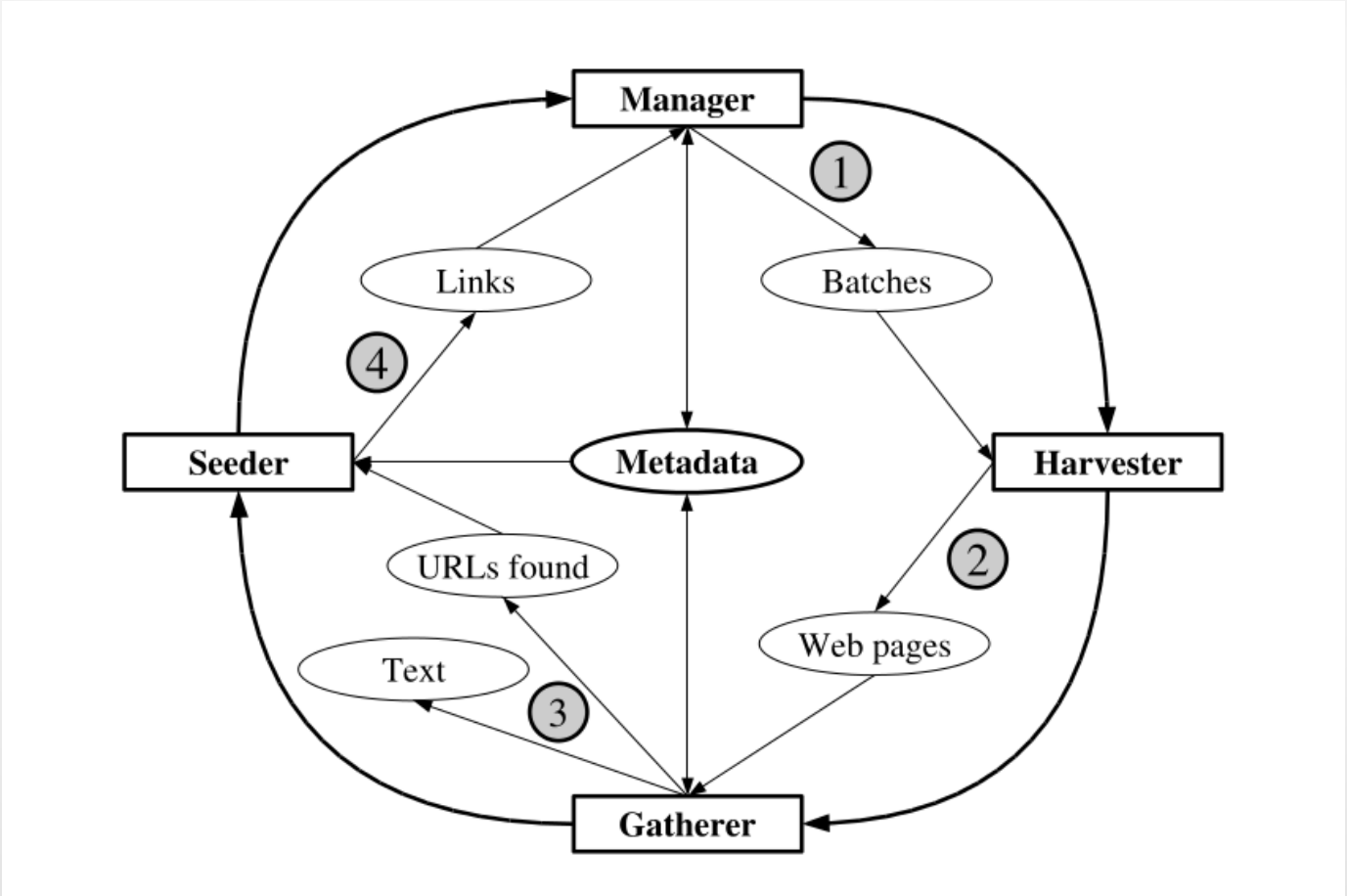
Modelagem em Grafo

- Relaxamento: dado um grafo G(V,A) e um inteiro *k*, em que:
  - os vértices de G são páginas da *Web,*
  - as arestas representam os *links* entre estas páginas, e
  - cada vértice possui uma pontuação de relevância não-conhecida *a priori,*

  percorrer um caminho que maximiza as pontuações dos vértices visitando no máximo *k* páginas.

Baseline

- **WIRE** (*Web Information Retrieval Environment*) [Castillo2004]



Heurística Utilizada

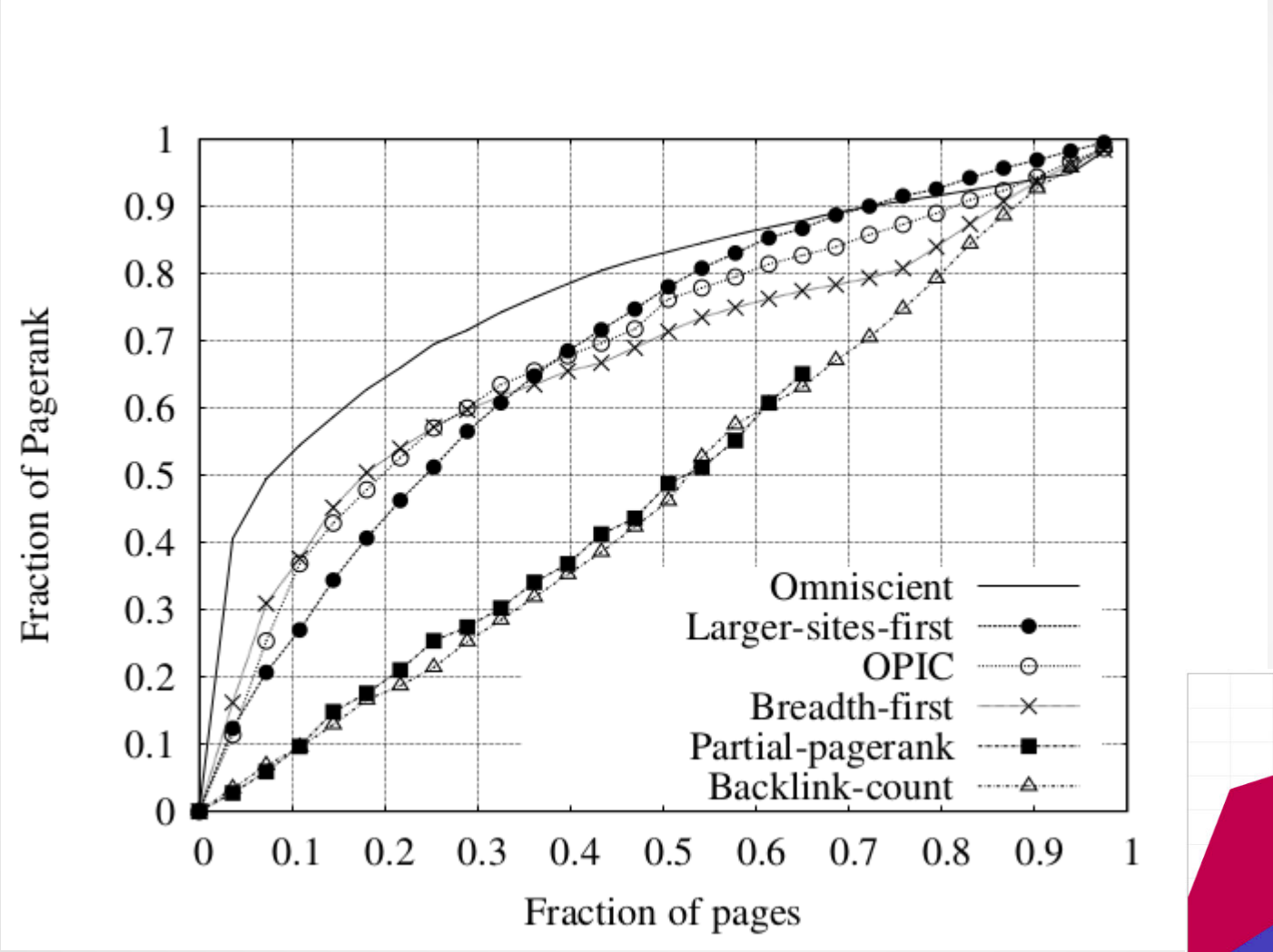
- Contagem de palavras (*Gatherer*)
- Análise de URL's e Pontuações (*Manager*)

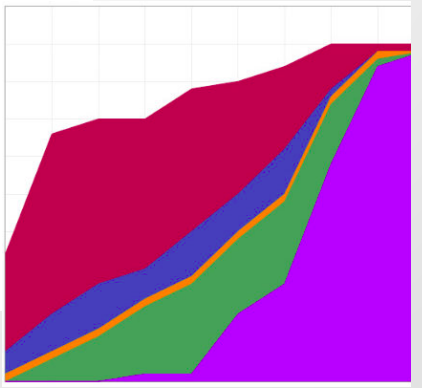
Análise Teórica de Complexidade

Algoritmo	**Complexidade de Tempo**	**Complexidade de Espaço**
Baseline (WIRE)	$O(V+E)$	$O(V+E)$
Heurística	$O(aV+bE) = O(V+E)$	$O(V+E+c) = O(V+E)$

- a, b, c: constantes em relação a V e E.

Experimentos e Resultados





Referências

- [Yates11] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval: The concepts and technology behind search, 2011.
- [Castillo04] Carlos Castillo. Effective Web Crawling. PhD thesis, School of Engineering, Santiago, Chile, November 2004.
- [Craswell04] Nick Craswell, Francis Crimmins, David Hawking, and Alistair Moffat. Performance and cost tradeoffs in web search, New Zealand, January 2004.
- [Gulli05] A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages, New York, USA, 2005. ACM.
- [Nielsen04] Statistics for traffic referred by search engines and navigation directories to useit, 2004.
- [Page98] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.