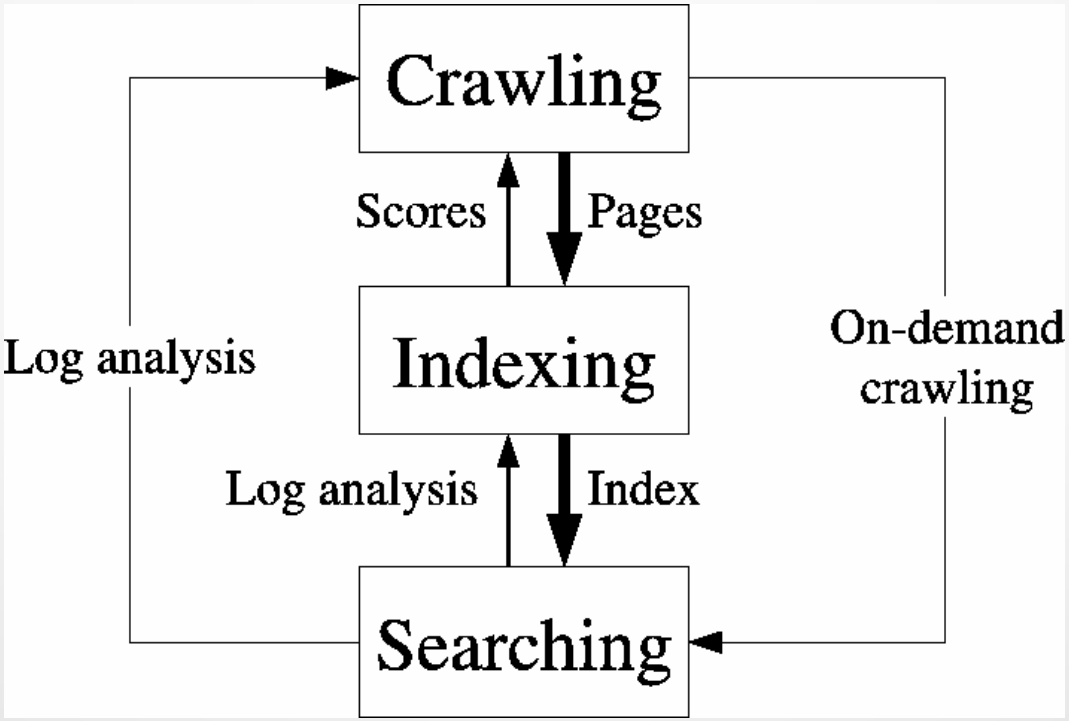
Uma Heurística para Focused Web Crawlers

Alberto Hideki Ueda
ueda@dcc.ufmg.br

O Problema

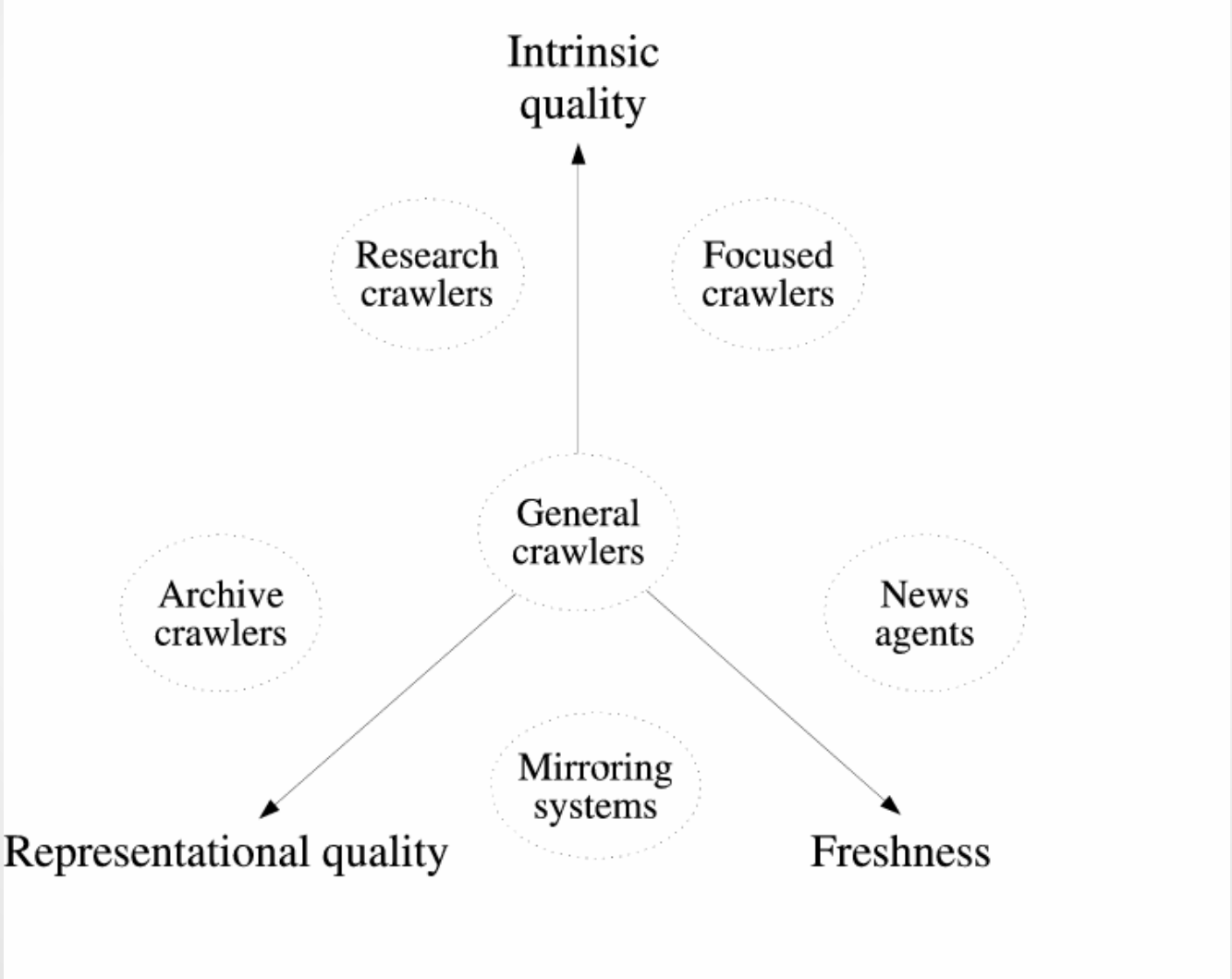
- Primeiro passo para a implementação de máquinas de busca modernas
- O que é um Coletor (*Crawler*)? [1]



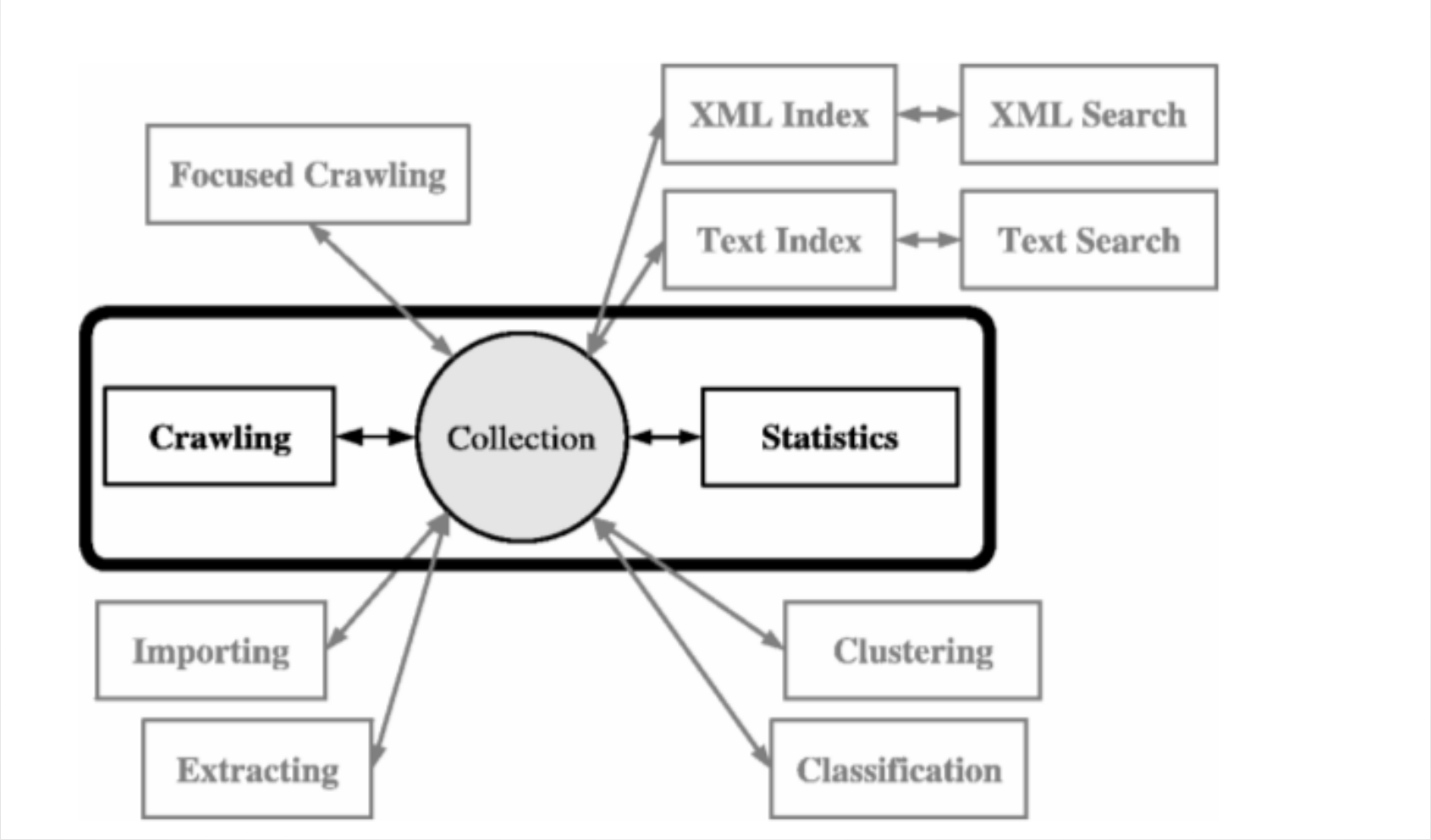
O Problema

- Primeiro passo para a implementação de máquinas de busca modernas
- O que é um Coletor (*Crawler*)? [1]
- Em 2005, foi demonstrado que o n´ıvel de cobertura das principais m´aquinas de buscas existentes est á entre 58% e 76% da Web [7].
- Em 2004, tal custo foi estimado em US$ 1,5 milhao para coletores de larga escala [6].
- **Inviabilidade da Solução Exata**

O Problema

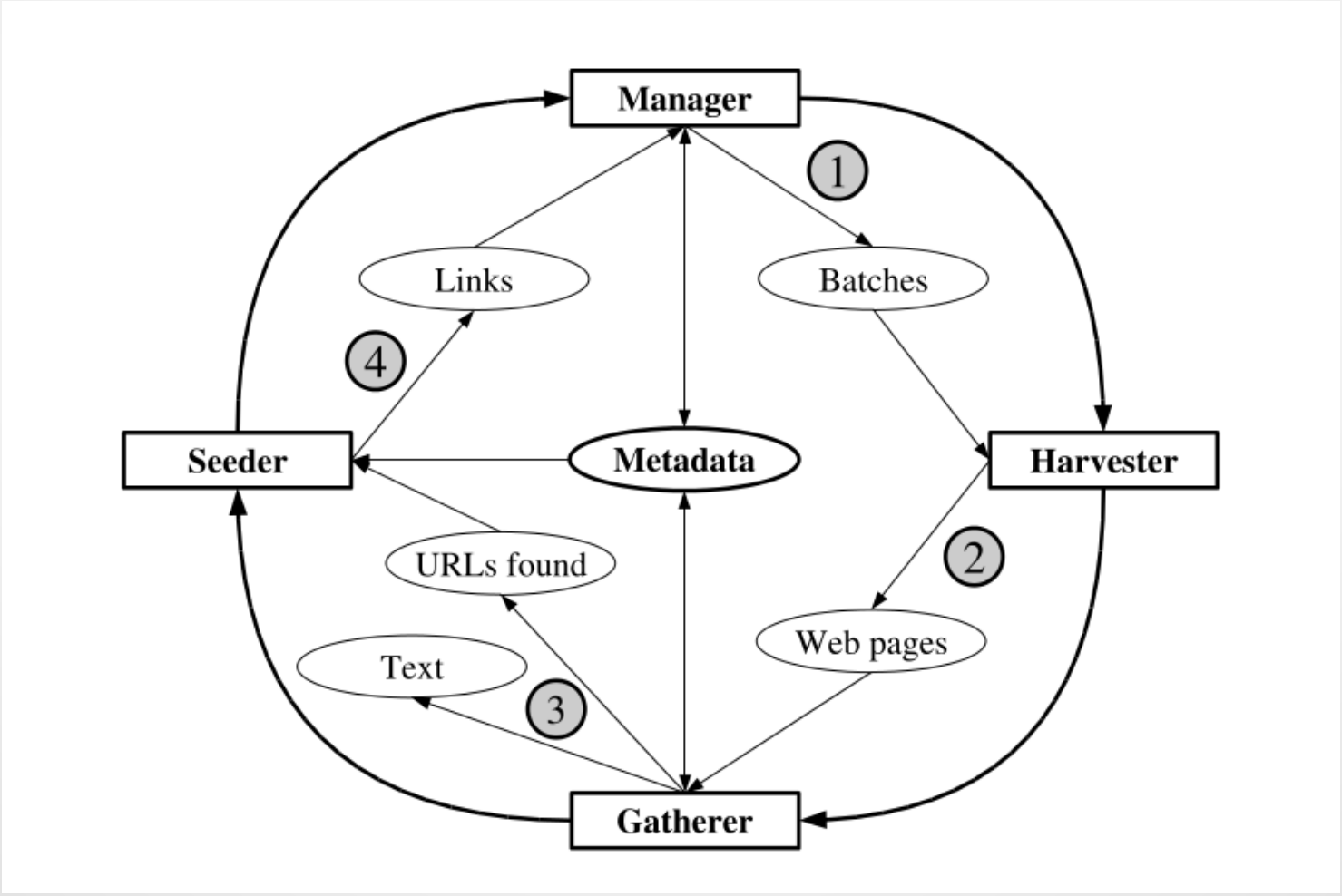

Intrinsic quality
Research crawlers
Focused crawlers
General crawlers
Archive crawlers
News agents
Mirroring systems
Representational quality
Freshness

O Problema

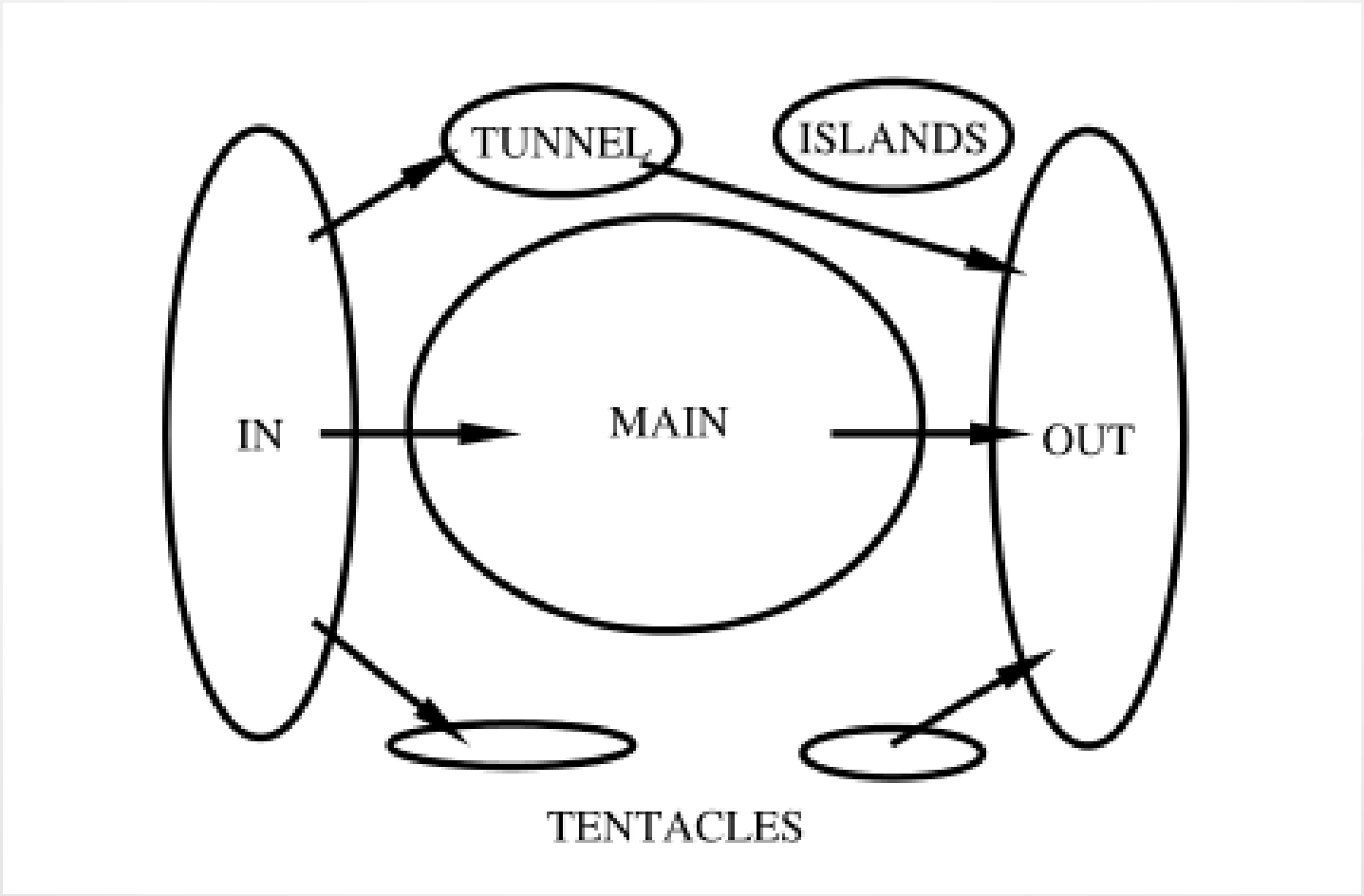

Focused Crawling
XML Index
XML Search
Text Index
Text Search
Crawling
Collection
Statistics
Importing
Extracting
Clustering
Classification

Baseline

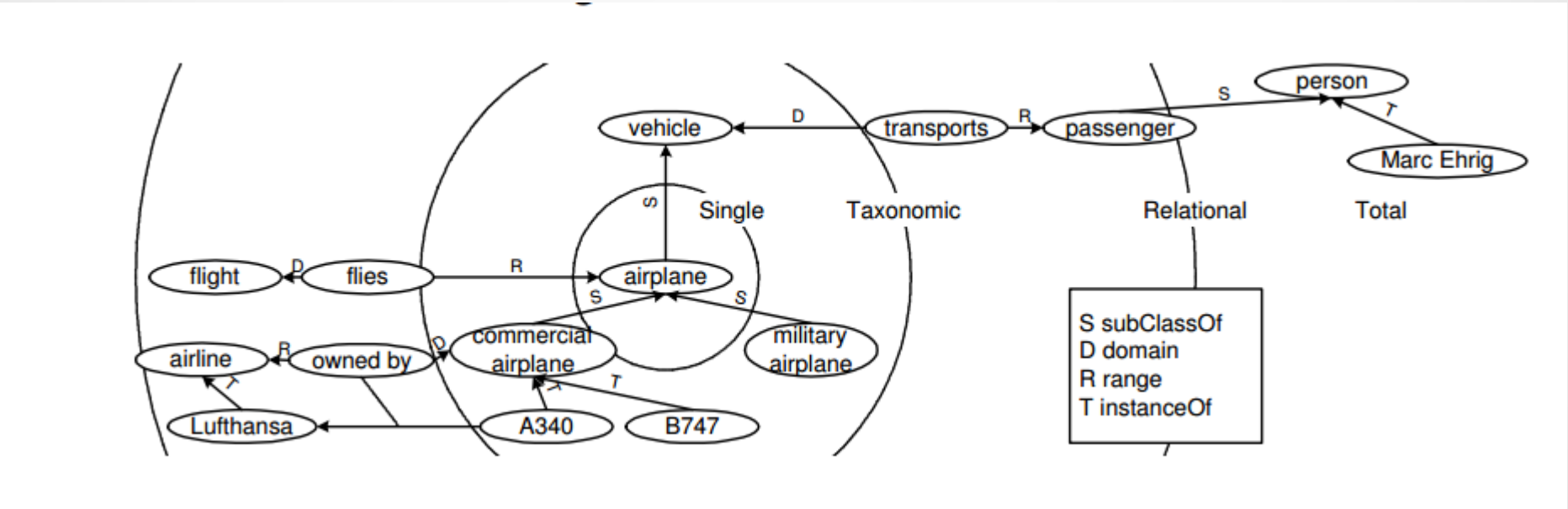
- **WIRE** [4]



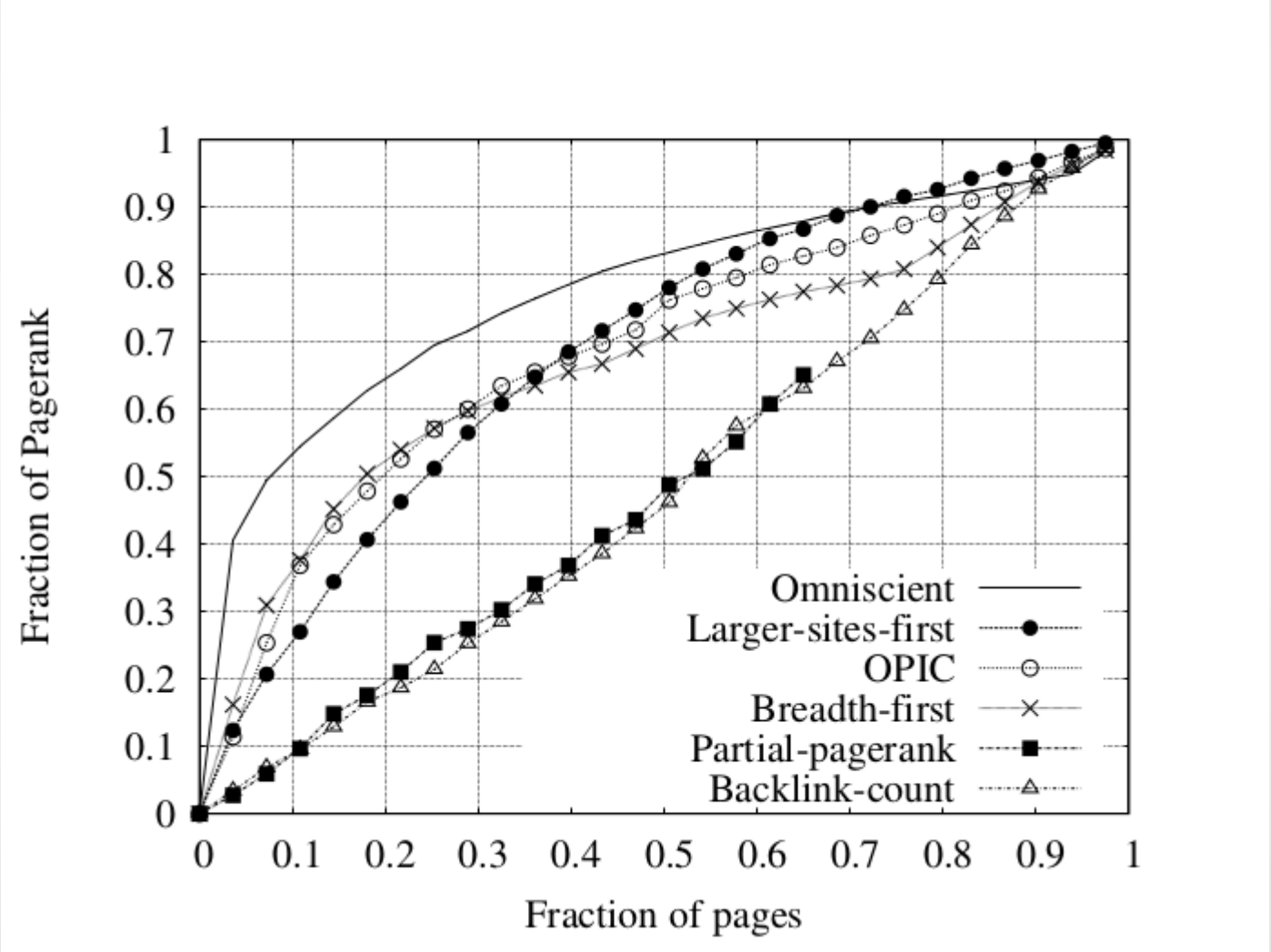
Modelagem em Grafo



Heurística Utilizada



Experimentos e Resultados



Referências

- Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval: The concepts and technology behind search. Pearson Education Limited, 2nd edition, 2011.

- Albert-Laszlo Barabasi. Linked the new science of networks, 2002.