UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Information Retrieval

Search Architecture

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br

Search infrastructure

Search engines run on resource-intensive regimes

- ***Bandwidth*** for handling crawling and search traffic
- ***Storage*** for persisting documents, indexes, metadata
- ***Processing*** for crawling, indexing and retrieval

Must scale from a single computer in one datacenter…

- … to huge clusters spread across availability zones

Financial costs

Depreciation

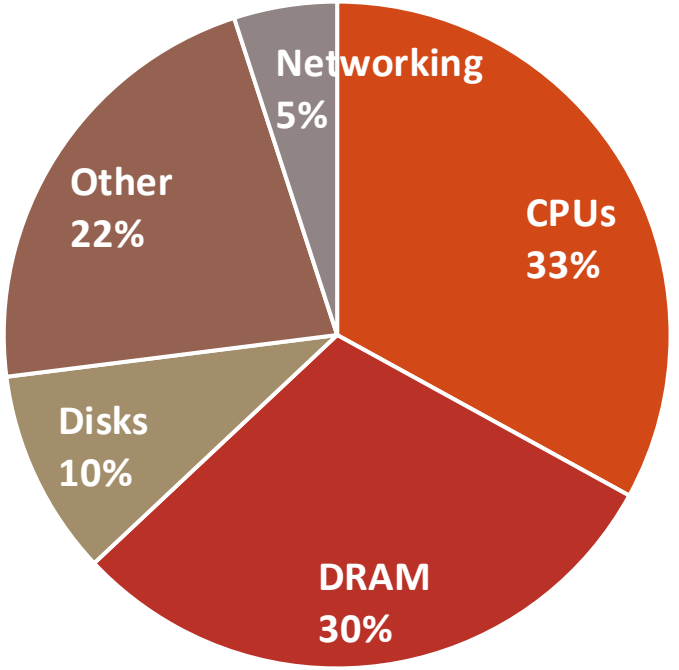
- Old hardware needs to be replaced

Maintenance

- Failures need to be handled

Operational

- Energy spending needs to be reduced



Credit: Jorge Jorquera

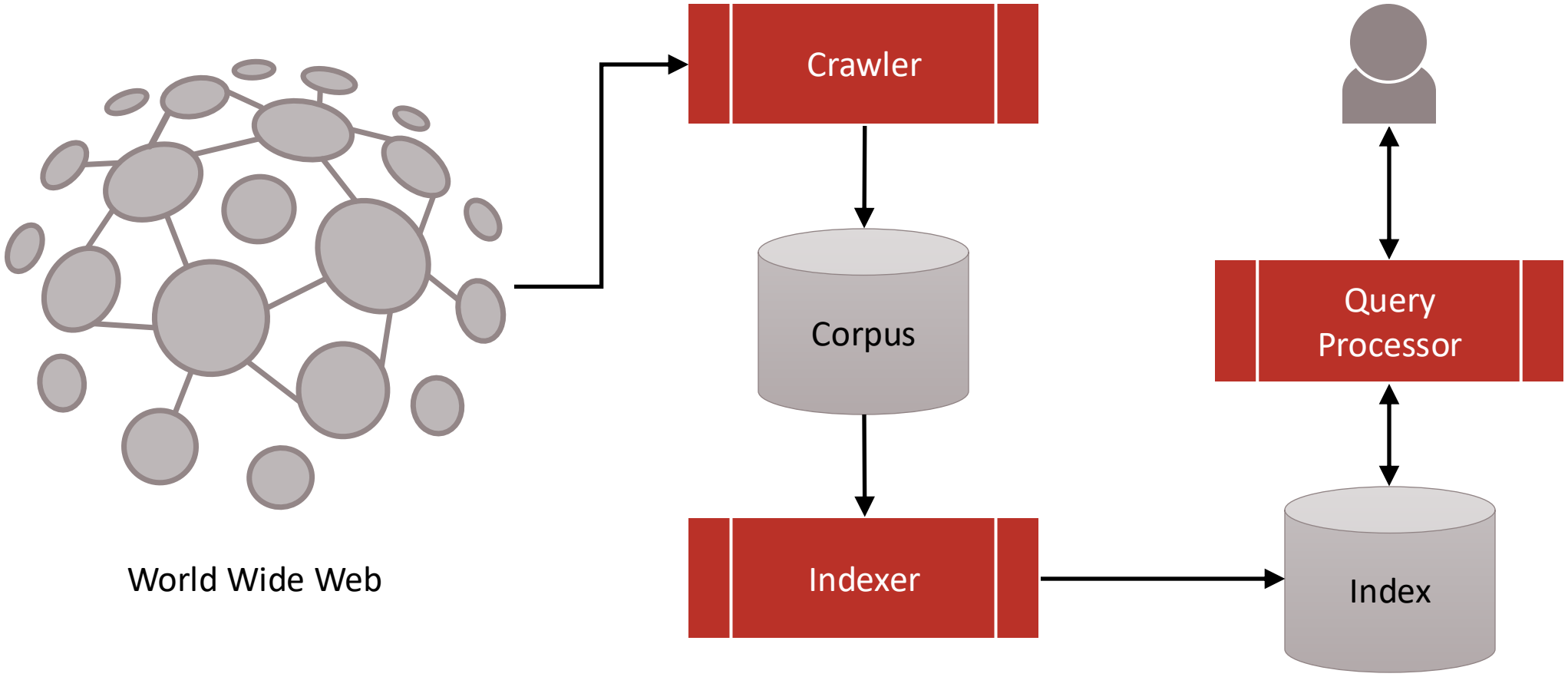
Search architecture

A software architecture consists of software components, the interfaces provided by those components, and the relationships between them

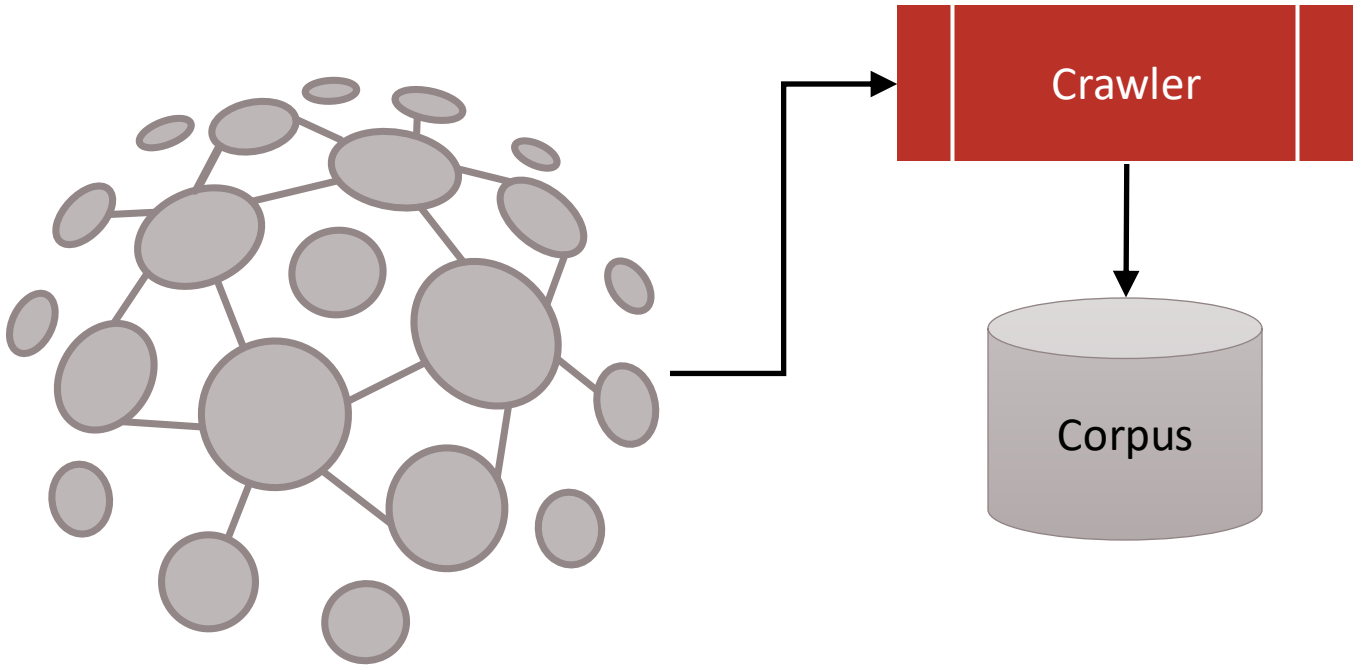
For search, we are concerned about

- Effectiveness (quality of results)
- Efficiency (response time and throughput)

Search components



Search components



World Wide Web

Crawling overview

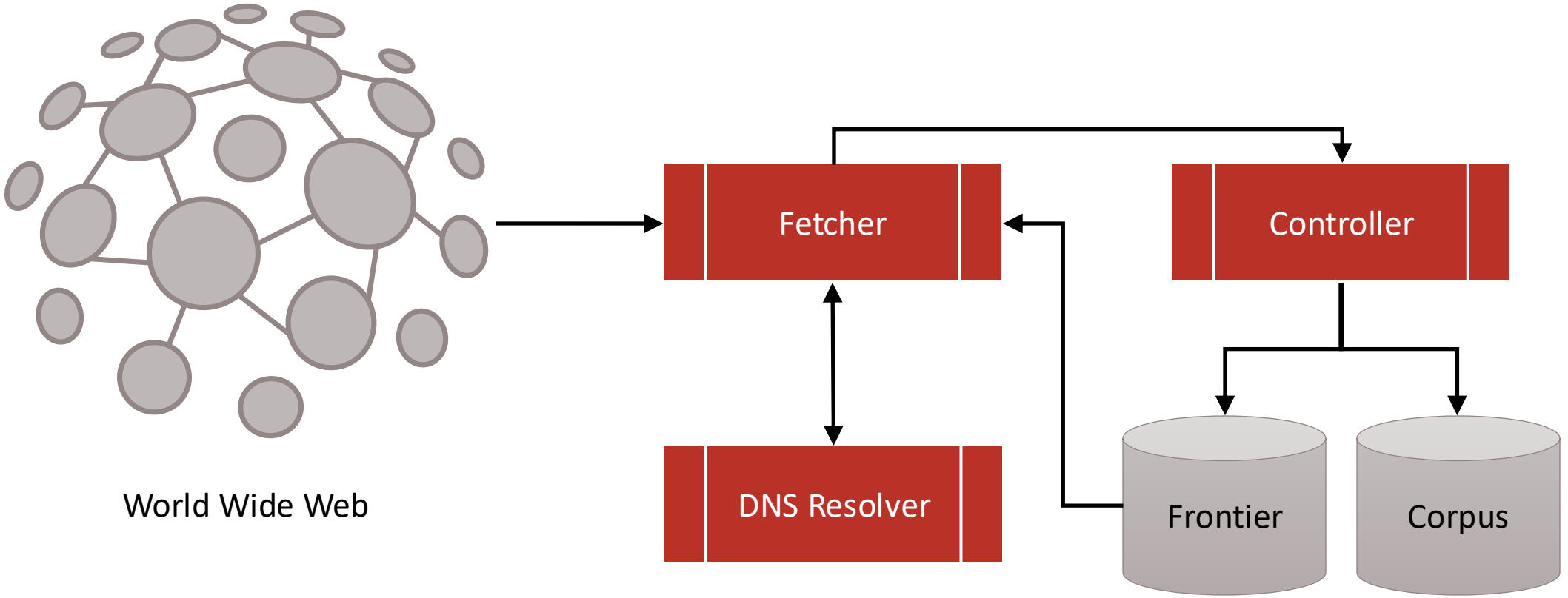
Document acquisition

- Builds a local corpus for searching
- Many types – Web, enterprise, desktop

Web crawlers follow links to find documents

- Must efficiently find huge numbers of web pages (coverage) and keep them up-to-date (freshness)

Crawling overview



Key challenges

Web is huge and constantly changing

- Not under the control of search providers

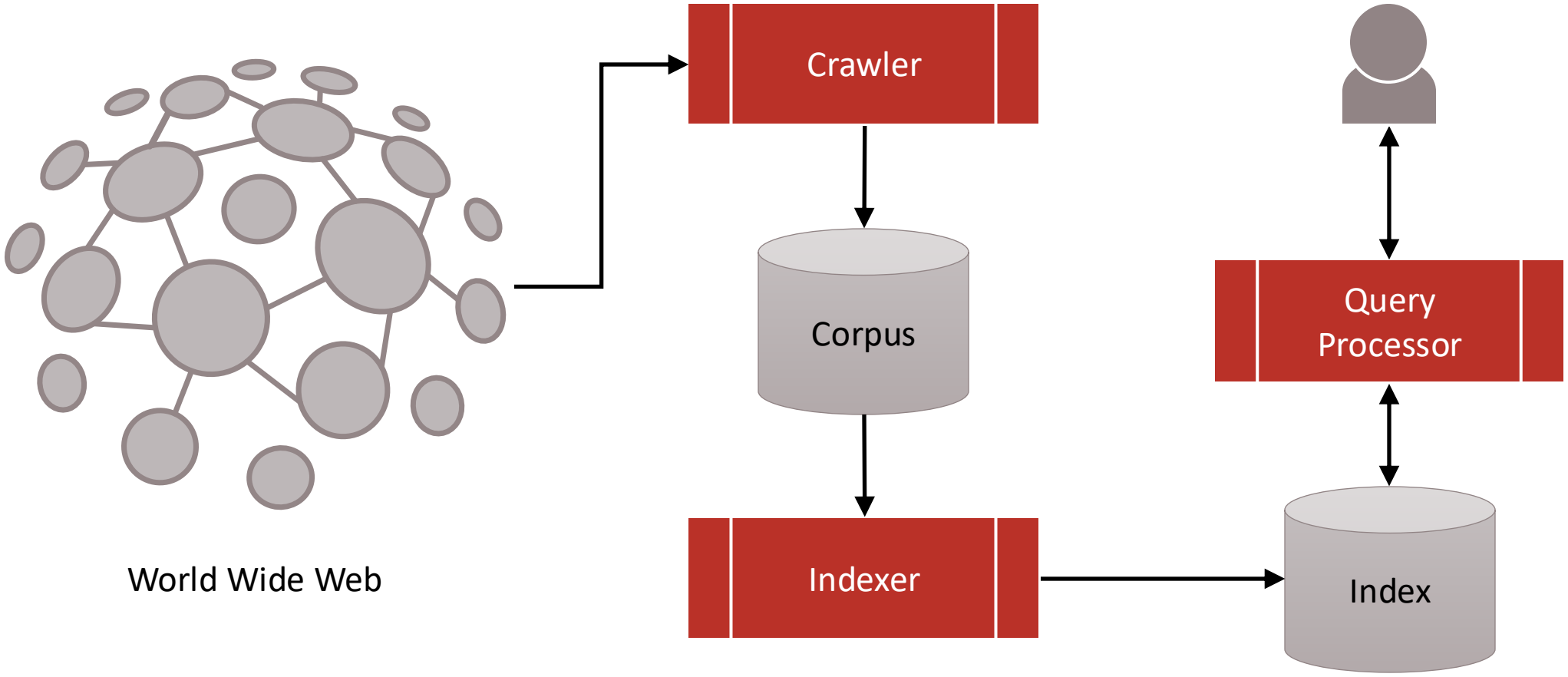
A lot of time is spent waiting for responses

- Parallel crawling is essential

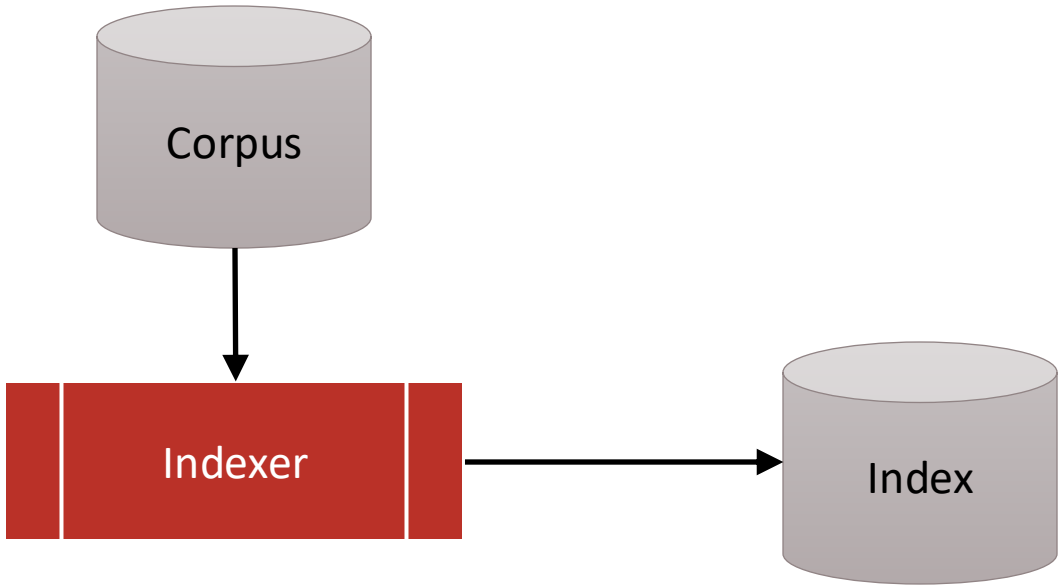
Could potentially flood sites with requests

- To avoid this problem, use politeness policies

Search components



Search components



Indexing overview

Document representation

- From raw text to index terms
- +annotations (e.g., entities, categories, embeddings)

Off-document evidence

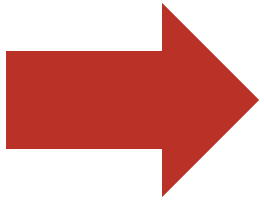
- Anchor text, link analysis
- Social network signals

Document representation

Fred’s Tropical Fish Shop is the best place to find tropical fish at low, low prices. Whether you’re looking for a little fish or a big fish, we’ve got what you need. We even have fake seaweed for your fishtank (and little surfboards too).

Document representation

Fred's Tropical Fish Shop is the best place to find tropical fish at low, low prices. Whether you're looking for a little fish or a big fish, we've got what you need. We even have fake seaweed for your fishtank (and little surfboards too).



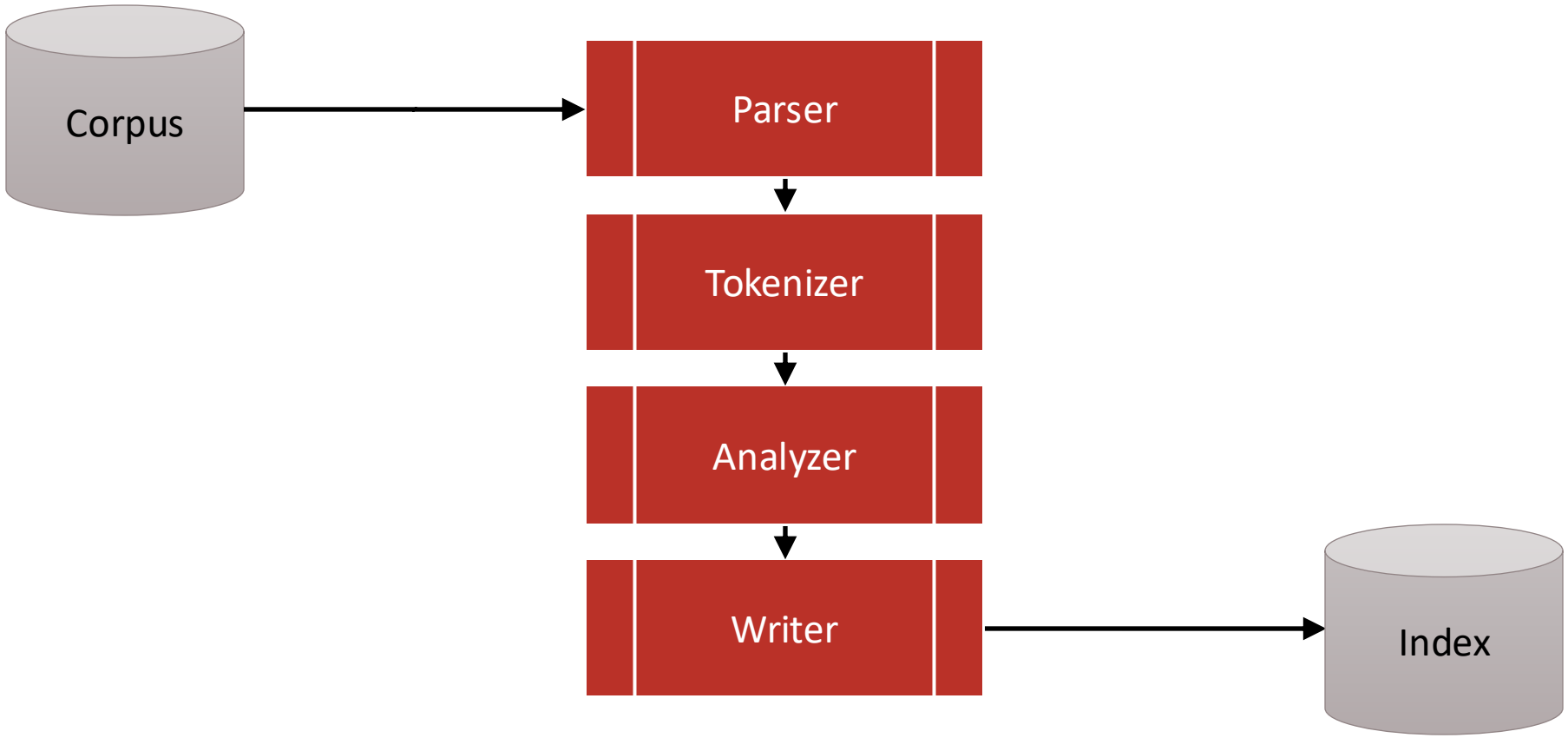
Topical features

9.7	fish
4.2	tropical
22.1	tropical fish
8.2	seaweed
4.2	surfboards

Quality features

14	incoming links
3	days since last update

Indexing overview



Key challenges

Support effective retrieval

- Extract meaningful document features
- Both topical and quality features

Support efficient retrieval

- Quick scoring of matched documents

Index structures

Indexes are designed to make search faster

- Unique requirements, unique data structures

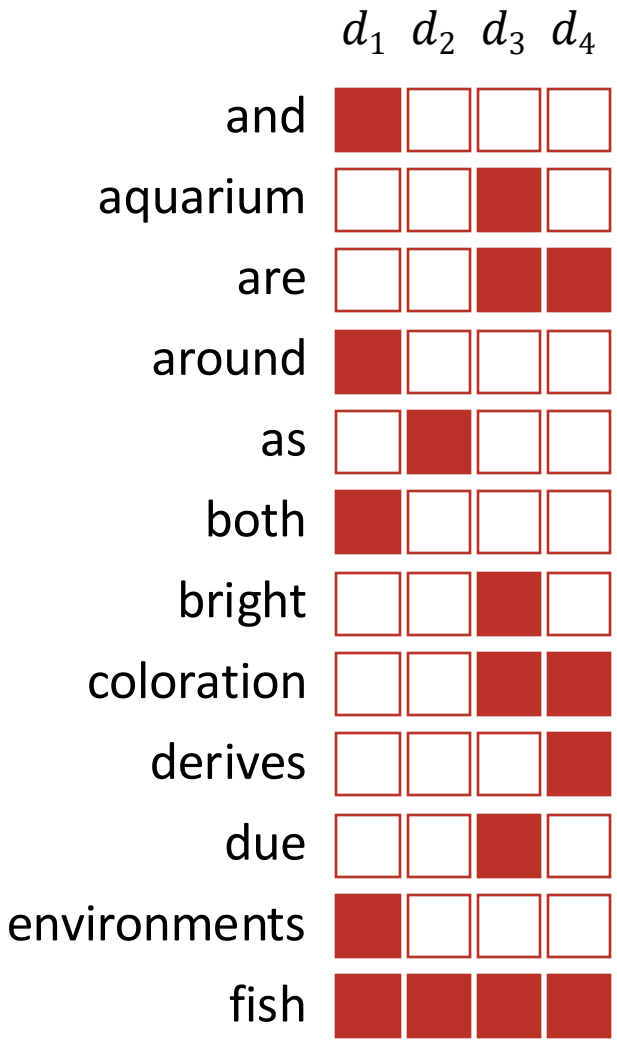
Most common structure is the inverted index

- General name for a class of structures
- “Inverted” because documents are associated with words, rather than words with documents

Example “corpus”

d_1	Tropical fish include fish found in tropical environments around the world, including both freshwater and salt water species.
d_2	Fish keepers often use the term tropical fish to refer only those requiring fresh water, with saltwater tropical fish referred to as marine fish.
d_3	Tropical fish are popular aquarium fish, due to their often bright coloration.
d_4	In freshwater fish, this coloration typically derives from iridescence, while salt water fish are generally pigmented.

Incidence matrix



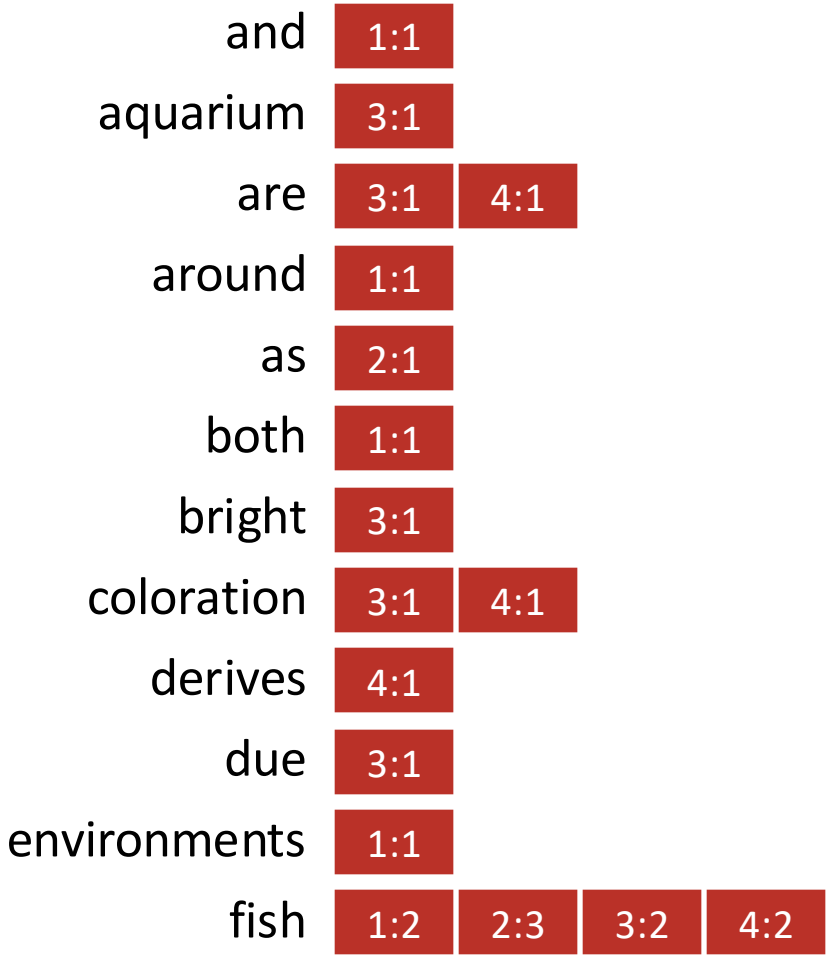
Straigthforward but…

… is it efficient?

Inverted index

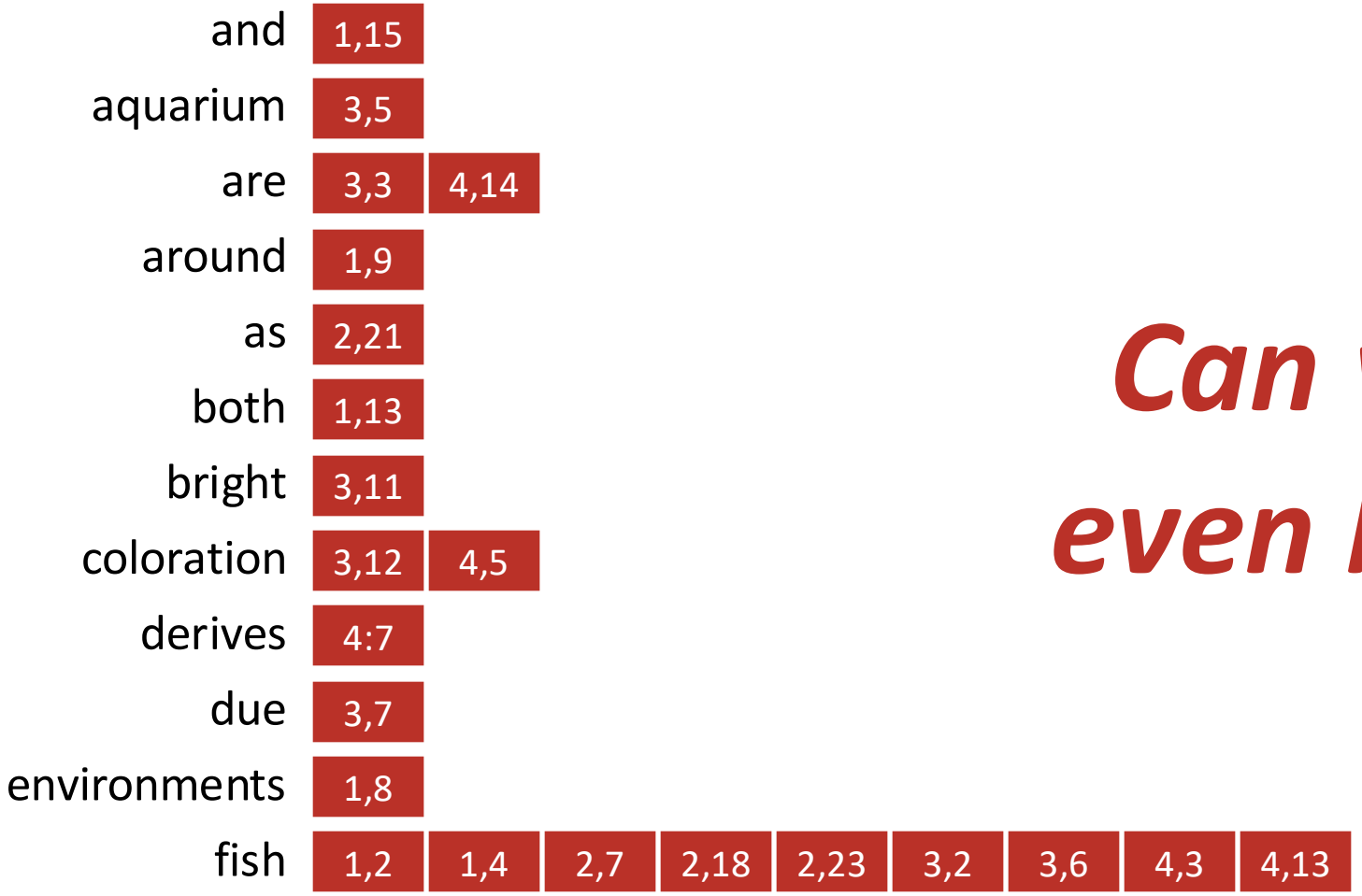
Aren't we missing anything?

Inverted index: counts



Can we do better?

Inverted index: positions



Can we do even better?

Inverted index: fields

Document structure is useful in search

- Field restrictions (e.g., date:, from:)
- Some fields more important (e.g., title, h1)

A couple of options

- Separate inverted lists for each field type
- Add information about fields to postings

Auxiliary structures

Vocabulary, dictionary, or lexicon

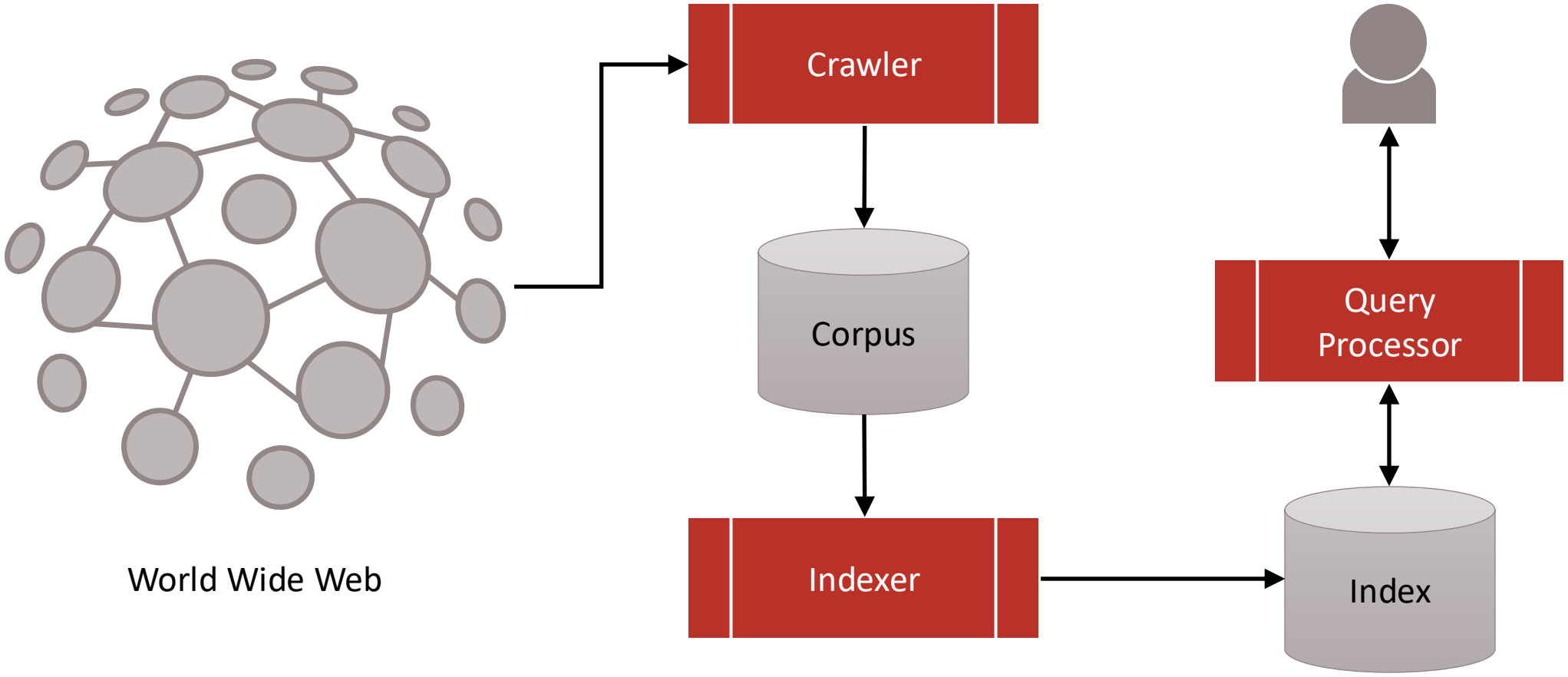
- Lookup table from term to inverted list
- Either hash table in memory or B-tree for disk

Additional structures for document data

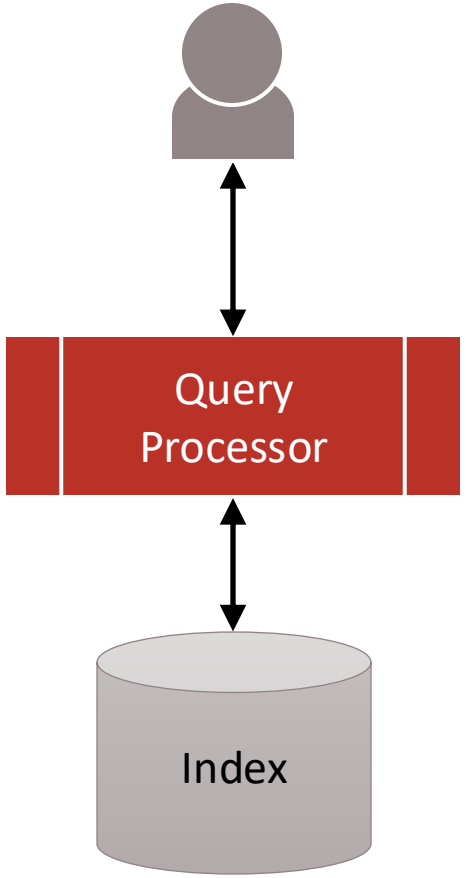
- Basic statistics, static features, metadata

Additional structure for corpus statistics

Search components



Search components



Query processing overview

Query representation

- Infers user's need from a keyword query

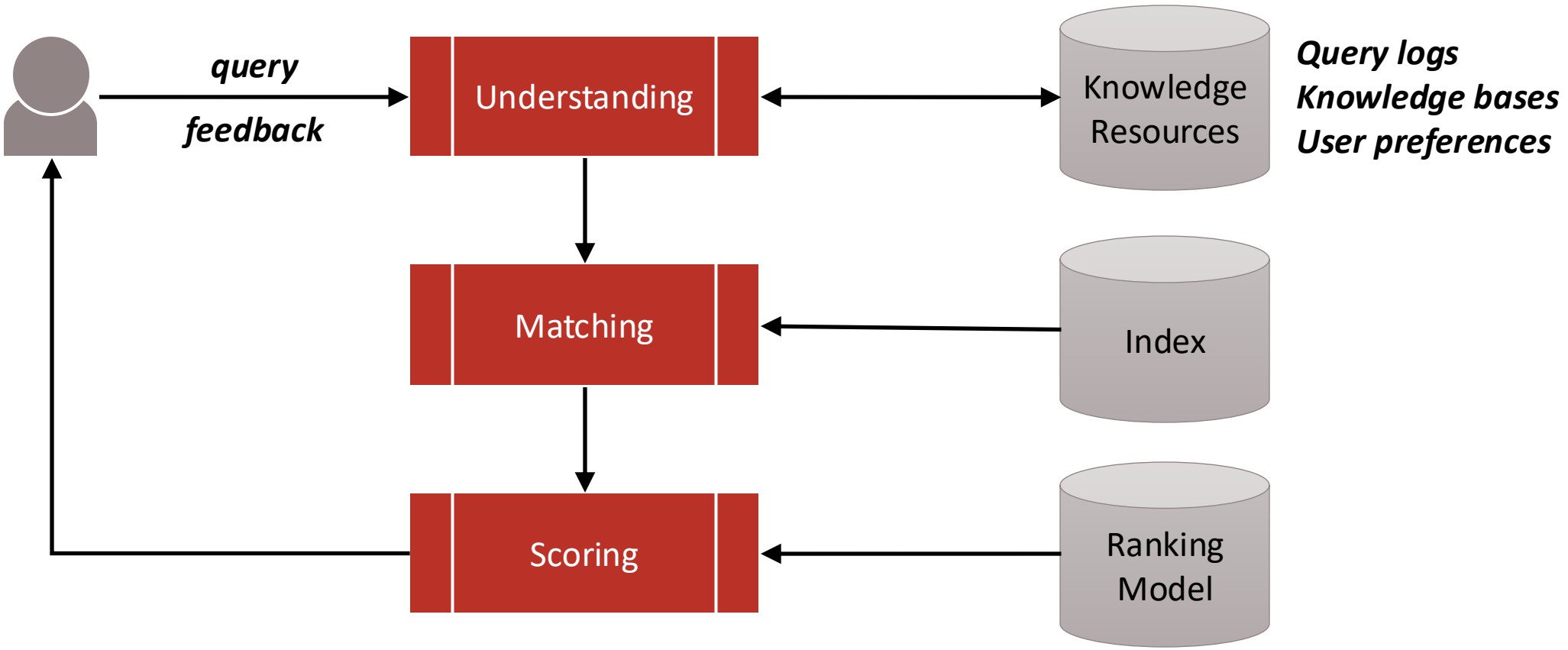
Document ranking

- Matches and scores indexed documents

Feedback handling

- Both explicit and implicit signals

Query processing overview



Key challenges

Queries are typically short, ill-specified

- Long queries tend to be difficult

Finding matching documents can be expensive

- Particularly for common terms or long queries

Ranking is a tough business

- Different queries, different requirements

Query understanding: expand matches

Query relaxation

- [information about tropical fish]

  ↳ [tropical fish]

Query expansion

- [tropical fish]

  ↳ [tropical fish aquarium]

Query understanding: narrow results

Query segmentation

- [tropical fish captive breeding]

  ↳ [”tropical fish” AND ”captive breeding”]

Query scoping

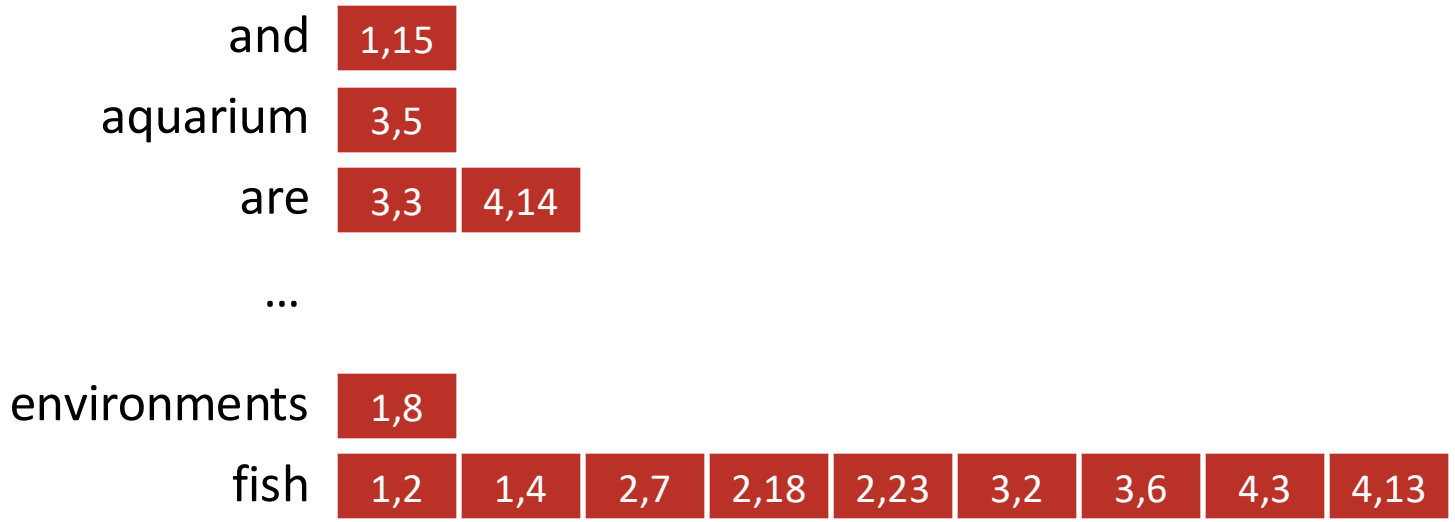
- [tropical fish hawaii]

  ↳ [category:”tropical fish” place:hawaii]

Document matching

Scan postings lists for all query terms

- [aquarium fish]



Document matching

Scan postings lists for all query terms

- [aquarium fish]

Score matching documents

- $f(q,d) = \sum_{t \in q} f(t,d)$

Document ranking

Many alternatives

- Lexical models (bag-of-words)
- Structural models (query + document structure)
- Semantic models (implicit + explicit semantics)
- Interactive models (user feedback)
- Feature-based models (aka learning to rank)

Summary

Search is a tough business

- Big data, big usage

An architecture tailored for efficiency is crucial

- Crawling, indexing, query processing

Must also cater for effectiveness

- Rule of thumb: don’t throw anything away

References

Search Engines: Information Retrieval in Practice, Ch. 2

Croft et al., 2009

Scalability Challenges in Web Search Engines

Cambazoglu and Baeza-Yates, 2015

UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

Coming next…

Web Crawling

Rodrygo L. T. Santos
rodrygo@dcc.ufmg.br