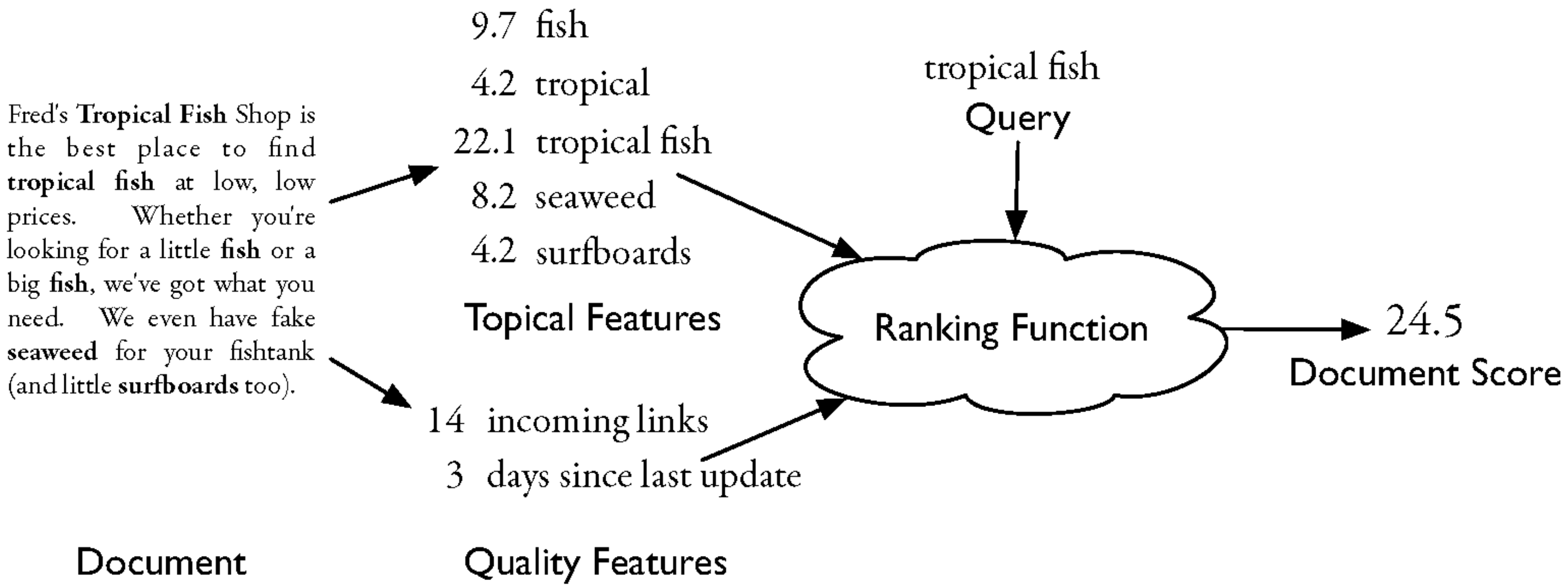
Search Engines

Information Retrieval in Practice

Abstract Model of Ranking



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

9.7 fish

4.2 tropical

22.1 tropical fish

8.2 seaweed

4.2 surfboards

Topical Features

14 incoming links

3 days since last update

Quality Features

tropical fish

Query

Ranking Function

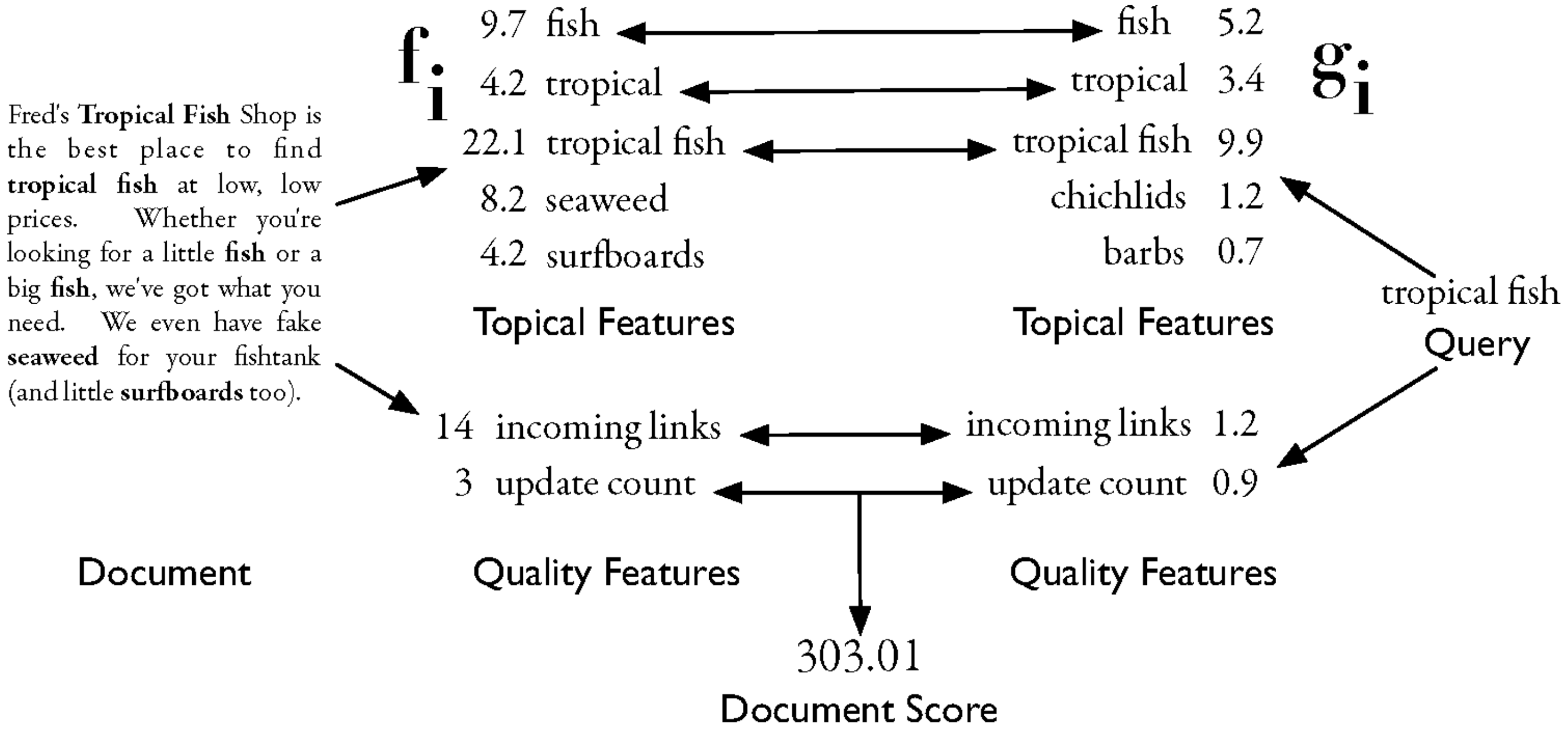
24.5

Document Score

More Concrete Model

$$R(Q,D) = \sum_i g_i(Q) f_i(D)$$

f_i is a document feature function
g_i is a query feature function



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

$\mathbf{f_i}$

9.7	fish
4.2	tropical
22.1	tropical fish
8.2	seaweed
4.2	surfboards

Topical Features

14	incoming links
3	update count

Quality Features

$\mathbf{g_i}$

fish	5.2
tropical	3.4
tropical fish	9.9
chichlids	1.2
barbs	0.7

Topical Features

incoming links	1.2
update count	0.9

Quality Features

tropical fish
Query

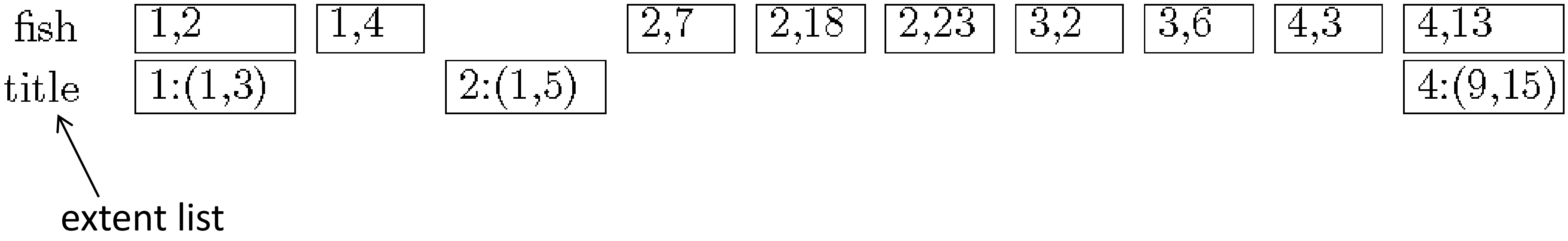
303.01
Document Score

Fields and Extents

- Document structure is useful in search
  - *field* restrictions
    - e.g., date, from:, etc.
  - some fields more important
    - e.g., title
- Options:
  - separate inverted lists for each field type
  - add information about fields to postings
  - use *extent lists*

Extent Lists

- An *extent* is a contiguous region of a document
  - represent extents using word positions
  - inverted list records all extents for a given field type
  - e.g.,



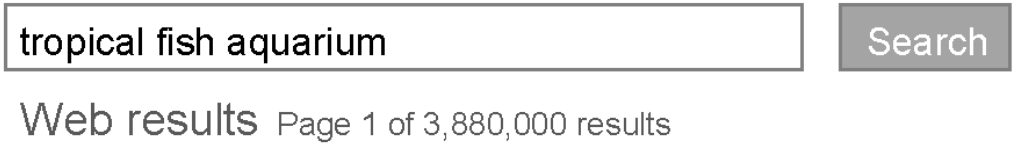
fish	1,2	1,4		2,7	2,18	2,23	3,2	3,6	4,3	4,13
title	1:(1,3)		2:(1,5)							4:(9,15)

extent list

Other Issues

- Precomputed scores in inverted list
  - e.g., list for “fish” [(1:3.6), (3:2.2)], where 3.6 is total feature value for document 1
  - improves speed but reduces flexibility
- Score-ordered lists
  - query processing engine can focus only on the top part of each inverted list, where the highest-scoring documents are recorded
  - very efficient for single-word queries

Estimating Result Set Size



tropical fish aquarium

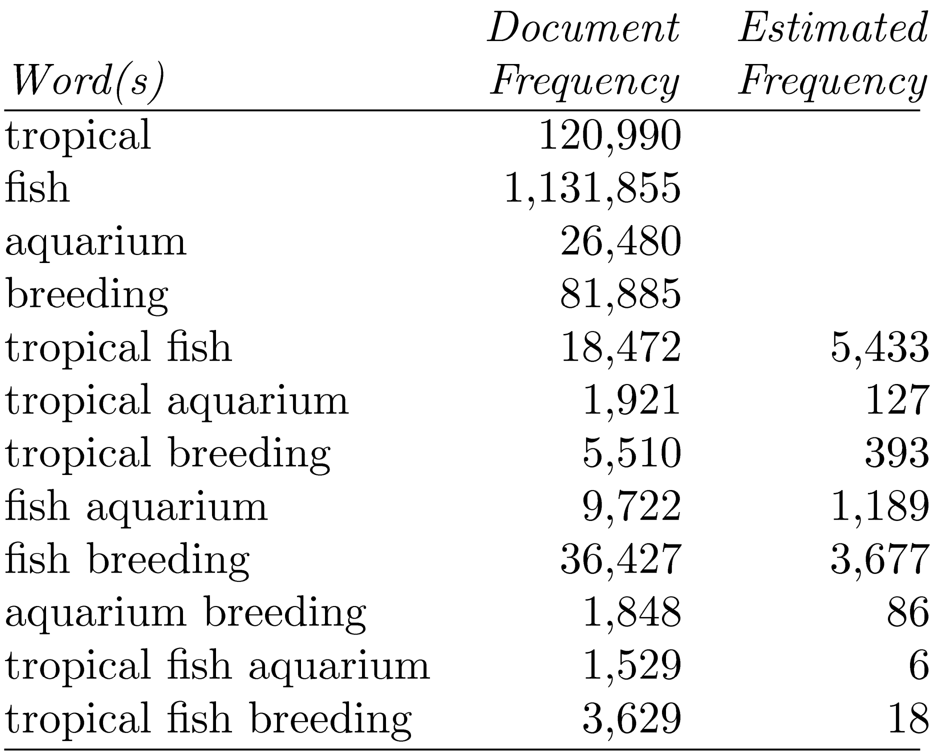
Web results Page 1 of 3,880,000 results

- How many pages contain *all* of the query terms?
- For the query *"a b c":*

$$f_{abc} = N \cdot f_a/N \cdot f_b/N \cdot f_c/N = (f_a \cdot f_b \cdot f_c)/N^2$$

  - Assuming that terms occur independently
  - f_{abc} is the estimated size of the result set
  - f_a, f_b, f_c are the number of documents that terms *a*, *b*, and *c* occur in
  - *N* is the number of documents in the collection

GOV2 Example



	Document	*Estimated*
Word(s)	*Frequency*	*Frequency*
tropical	120,990	
fish	1,131,855	
aquarium	26,480	
breeding	81,885	
tropical fish	18,472	5,433
tropical aquarium	1,921	127
tropical breeding	5,510	393
fish aquarium	9,722	1,189
fish breeding	36,427	3,677
aquarium breeding	1,848	86
tropical fish aquarium	1,529	6
tropical fish breeding	3,629	18

Collection size (N) is 25,205,179

Result Set Size Estimation

- Poor estimates because words are not independent
- Better estimates possible if co-occurrence information available

$$\mathrm{P}(a \cap b \cap c) = \mathrm{P}(a \cap b) \cdot \mathrm{P}(c|(a \cap b))$$

$$f_{tropical \cap fish \cap aquarium} = f_{tropical \cap aquarium} \cdot f_{fish \cap aquarium} / f_{aquarium}$$

$$= 1921 \cdot 9722/26480 = 705$$

$$f_{tropical \cap fish \cap breeding} = f_{tropical \cap breeding} \cdot f_{fish \cap breeeding} / f_{breeding}$$

$$= 5510 \cdot 36427/81885 = 2451$$

Result Set Estimation

- Even better estimates using initial result set
  - Estimate is simply *C*/*s*
    - where *s* is the proportion of the total documents that have been ranked, and *C* is the number of documents found that contain all the query words
  - E.g., “tropical fish aquarium” in GOV2
    - after processing 3,000 out of the 26,480 documents that contain “aquarium”, *C* = 258

$f_{tropical \cap fish \cap aquarium}$ = 258/(3000÷26480) = 2,277

    - After processing 20% of the documents,

$f_{tropical \cap fish \cap aquarium}$ = 1,778   (1,529 is real value)

Estimating Collection Size

- Important issue for Web search engines
- Simple technique: use independence model
  - Given two words *a* and *b* that are independent

    $f_{ab}/N = f_a/N \cdot f_b/N$

    $N = (f_a \cdot f_b)/f_{ab}$

  - e.g., for GOV2

    $f_{lincoln}$ = 771,326 $f_{tropical}$ = 120,990 $f_{lincoln \cap tropical}$ = 3,018

    N = (120990 · 771326)/3018 = 30,922,045

    (actual number is 25,205,179)