Uniform Split

PostingFormat

Contributors
Bruno Roustant, Juan Camilo Rodriguez Duran, David Smiley

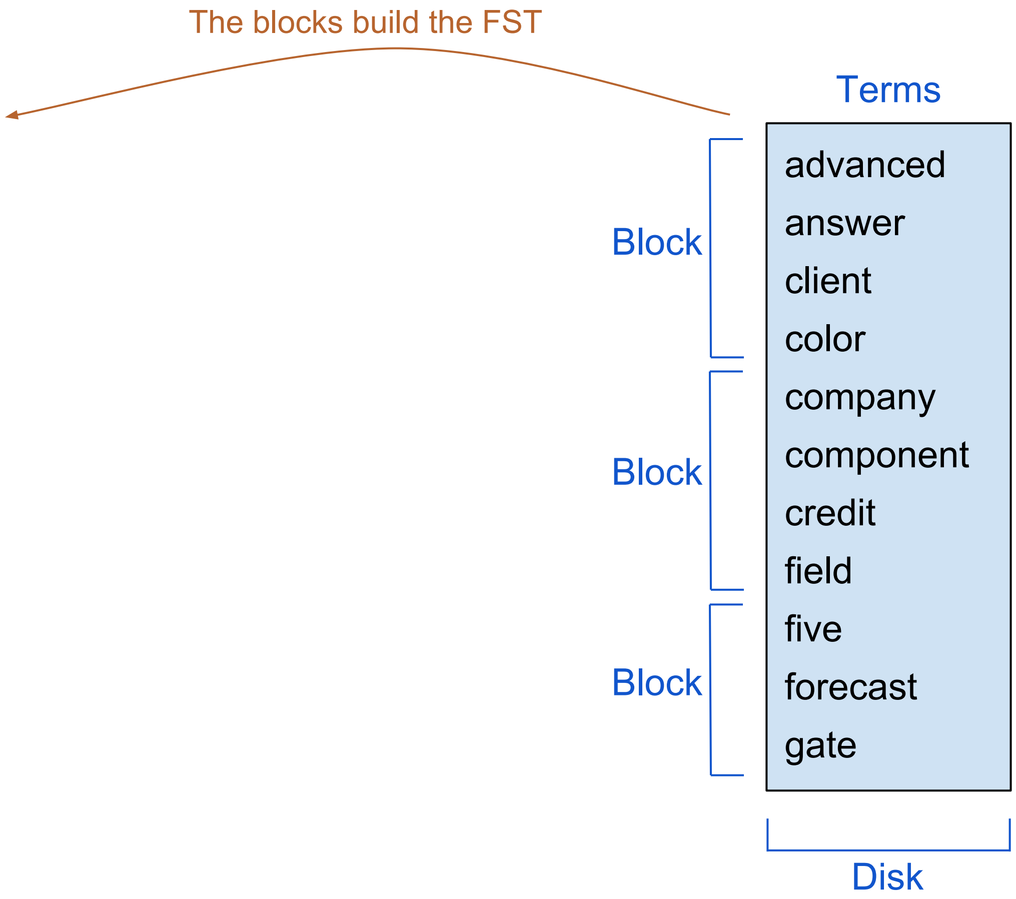
March, 2019

Building the dictionary

The blocks of uniform size build the FST

Good size = 32

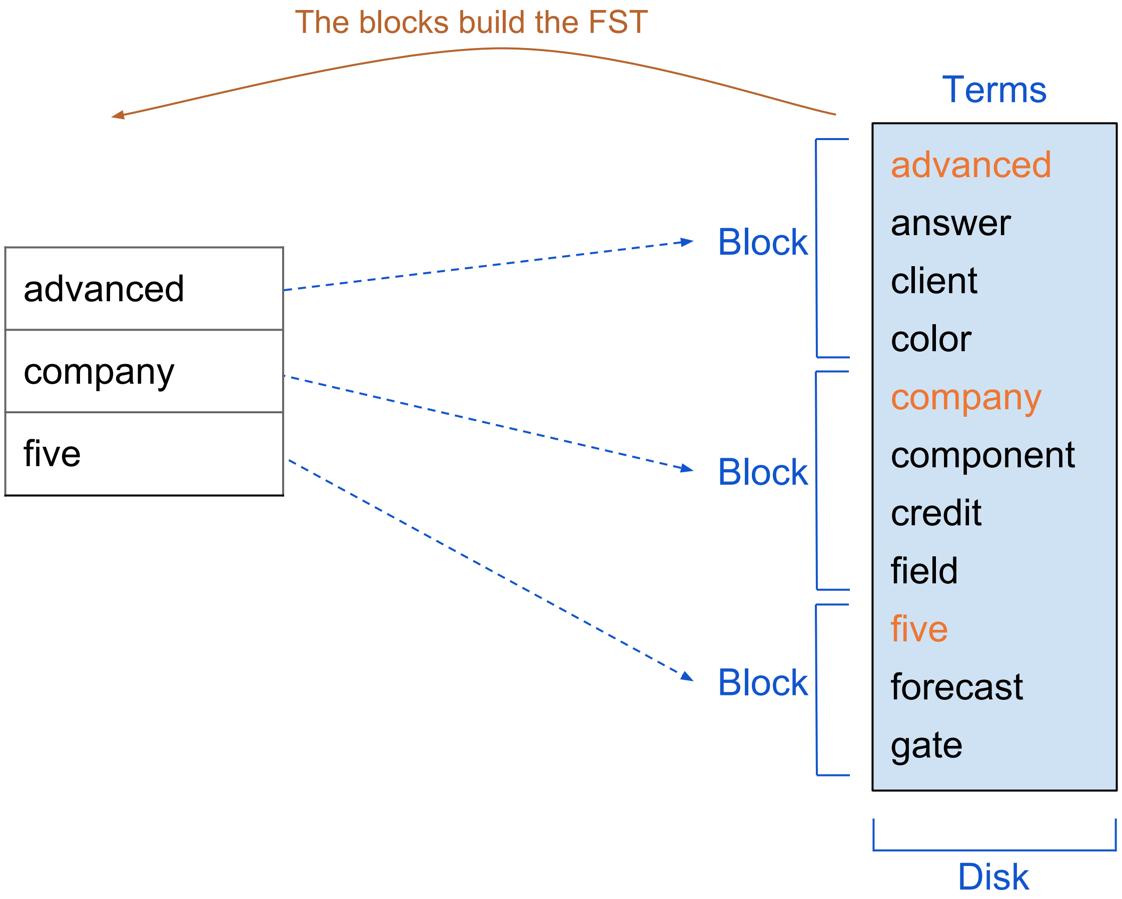
Size = 4 in this example



advanced
answer
client
color
company
component
credit
field
five
forecast
gate

Building the dictionary

The blocks of uniform size build the FST



The blocks build the FST

advanced

company

five

Block

Block

Block

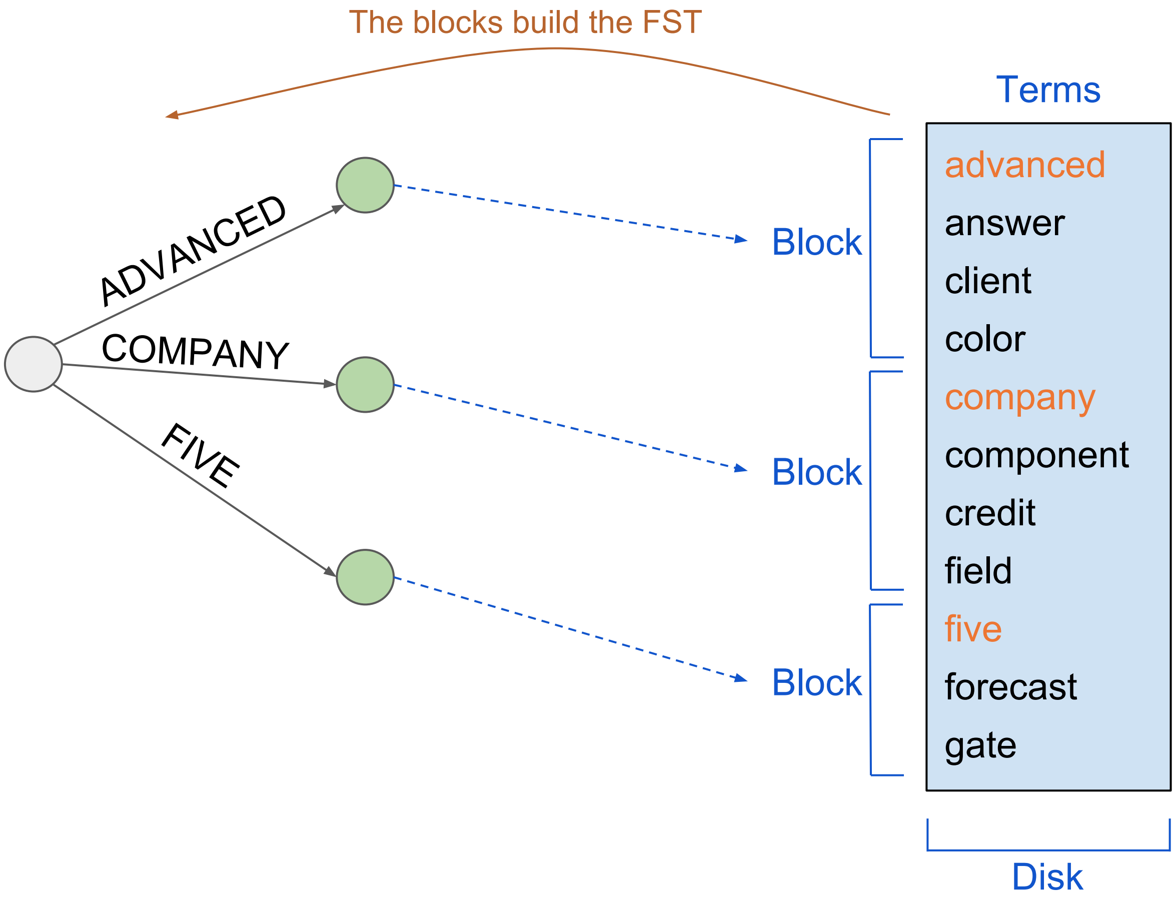
Terms

advanced
answer
client
color
company
component
credit
field
five
forecast
gate

Disk

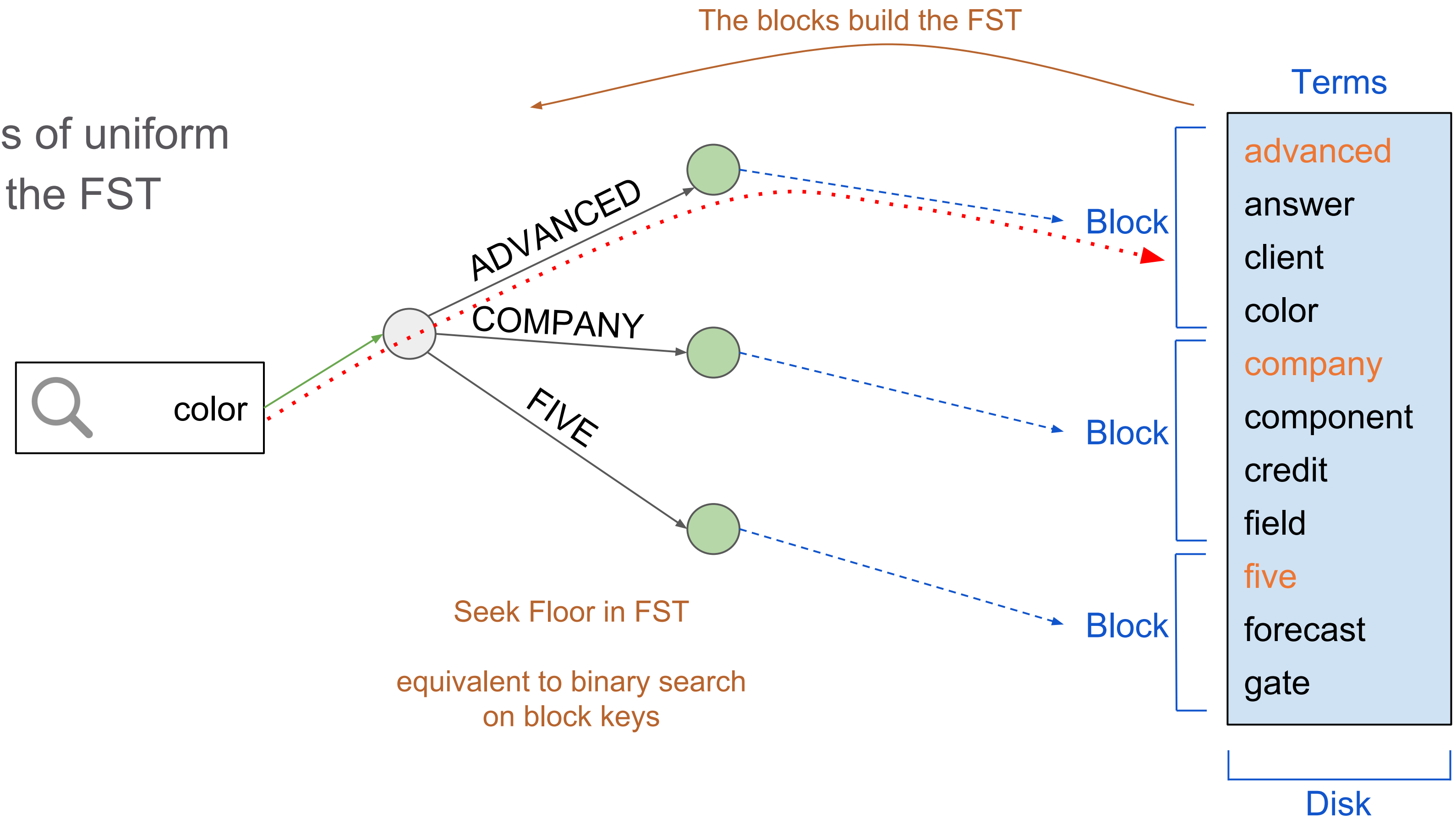
Building the dictionary

The blocks of uniform size build the FST



Building the dictionary

The blocks of uniform size build the FST



Seek Floor in FST

equivalent to binary search on block keys

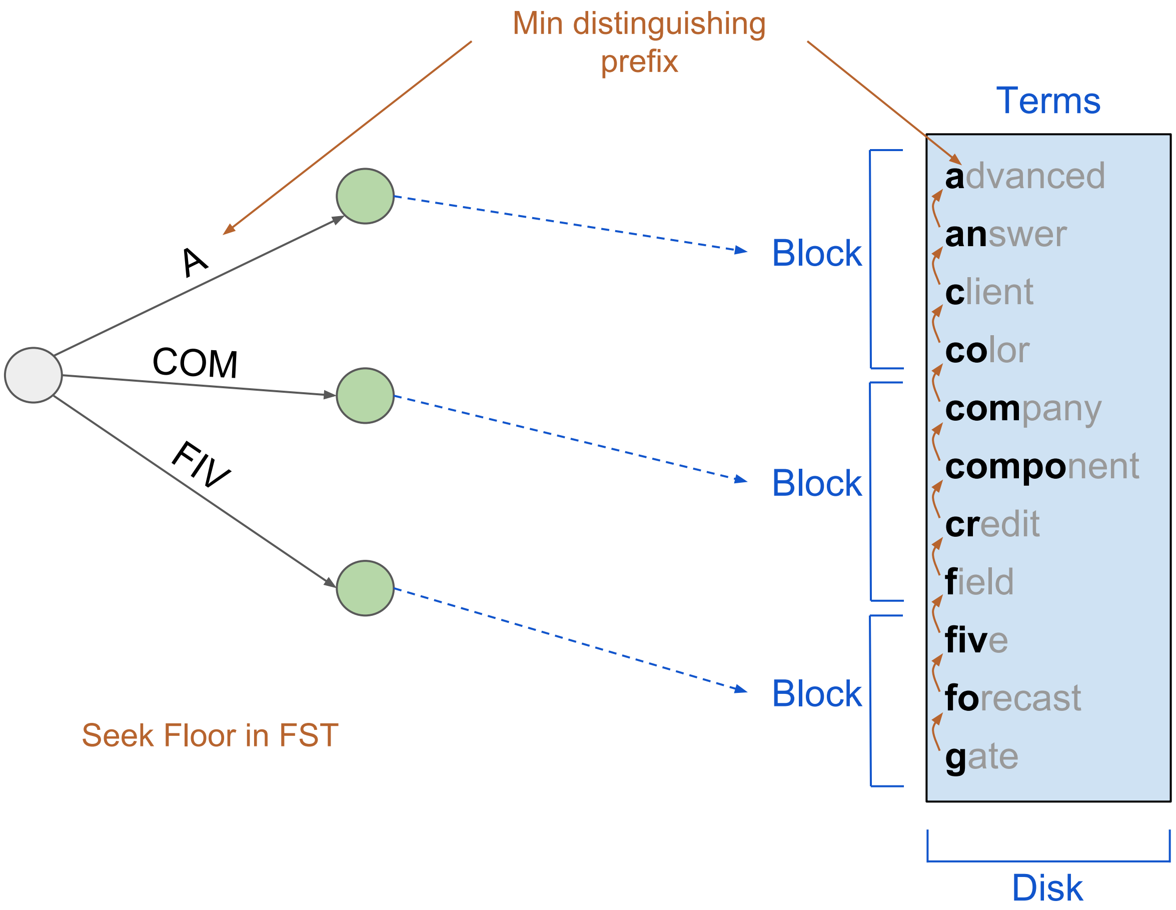
Building the dictionary

Optimization 1

Compute the **minimal distinguishing prefix (mdp)** for each term

More compact

Same search



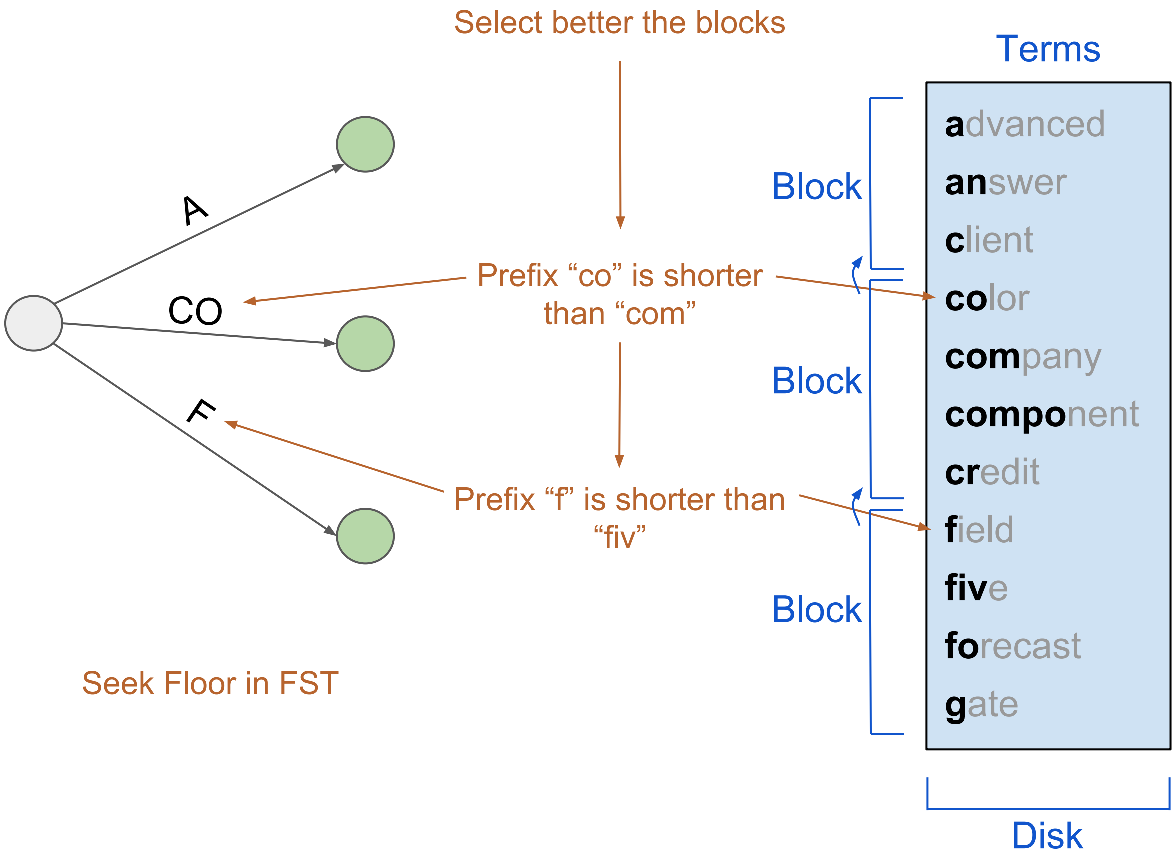
Building the dictionary

Optimization 2

Delta select the best minimal distinguishing prefix (mdp) per block Target block size +- 10%

More compact

Same search



Select better the blocks

Prefix “co” is shorter than “com”

Prefix “f” is shorter than “fiv”

Seek Floor in FST

Terms

advanced
answer
client
color
company
component
credit
field
five
forecast
gate

Disk

Building the dictionary

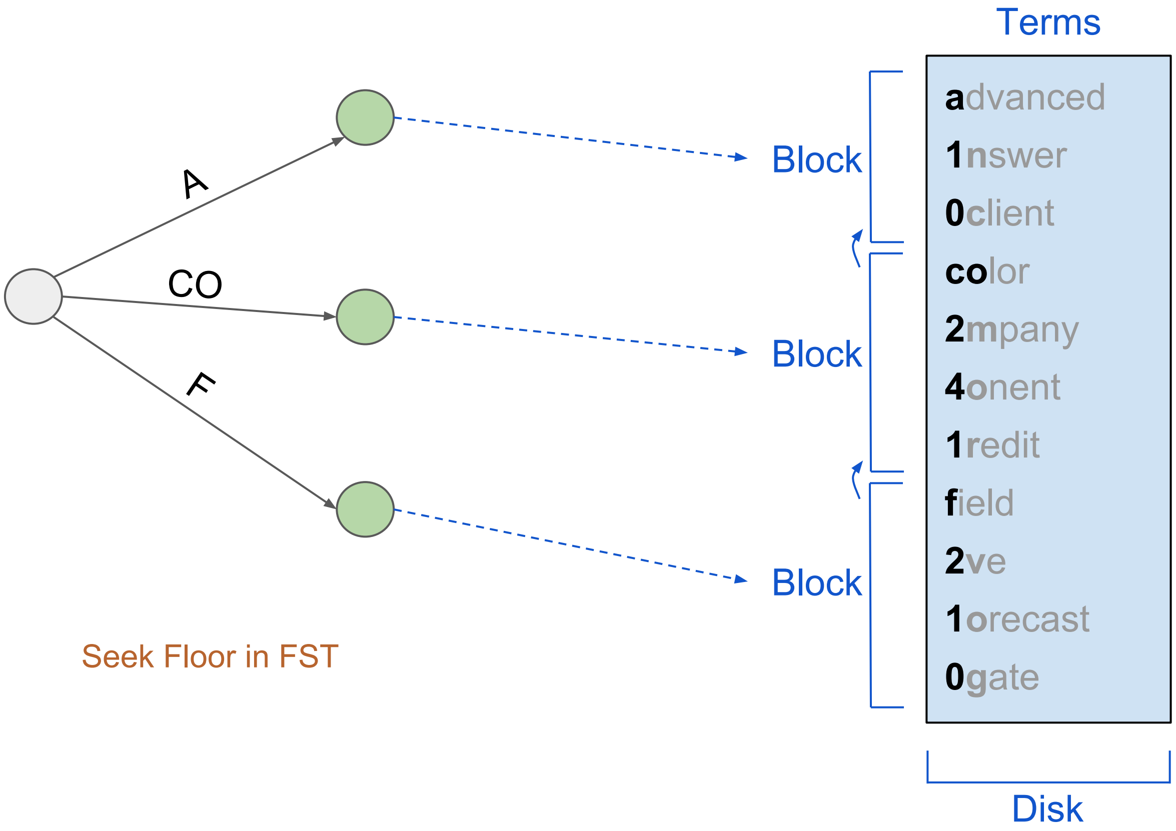
Optimization 3

Incrementally encode terms in each block

- Encode mdp length
- Only write suffix bytes

More compact

Faster scan



advanced
1nswer
0client
color
2mpany
4onent
1redit
field
2ve
1orecast
0gate

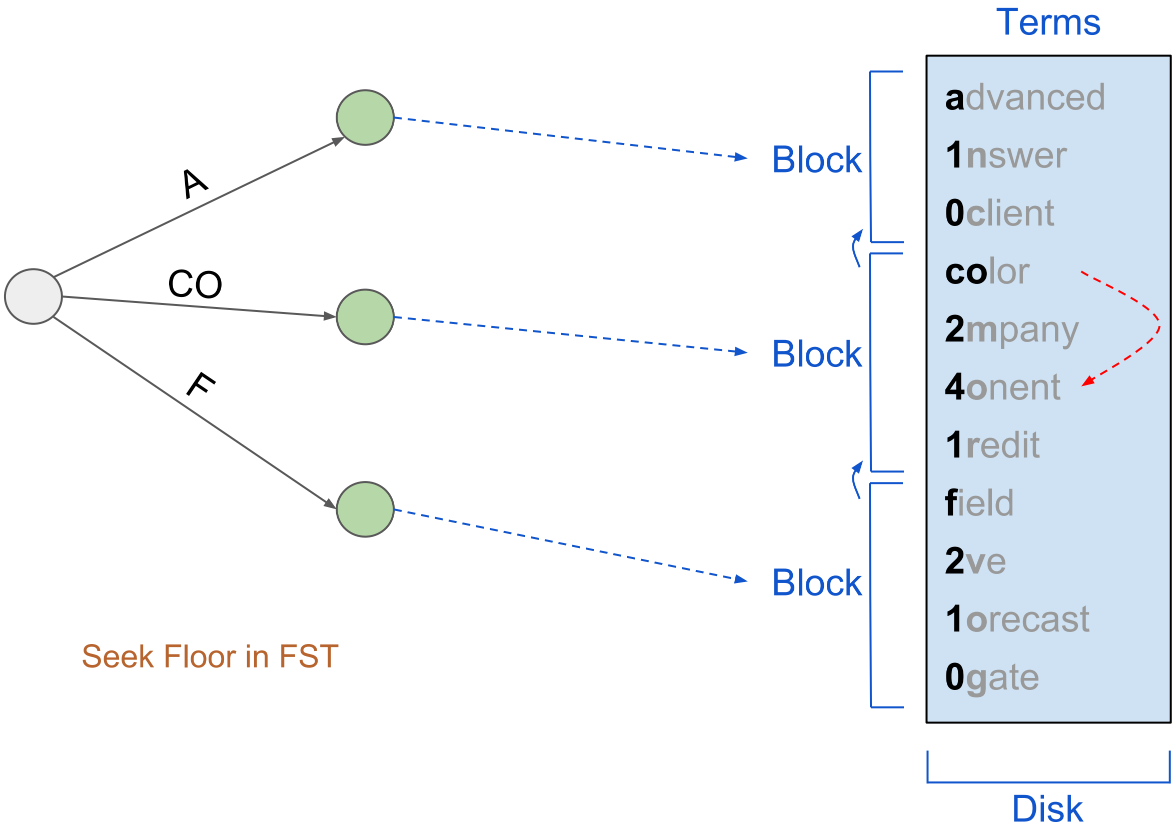
Seek Floor in FST

Building the dictionary

Optimization 4

Compare-&-jump to middle term inside a block
- Cheaply reduces by half the block scan

Faster scan


Terms
A
CO
F
Block
Block
Block
advanced
1nswer
0client
color
2mpany
4onent
1redit
field
2ve
1orecast
0gate
Seek Floor in FST
Disk