Document Retrieval on General Sequence Collections

Gonzalo Navarro
www.dcc.uchile.cl/gnavarro
gnavarro@dcc.uchile.cl

Department of Computer Science
University of Chile

Why this Problem?

Document Listing

Top-k Retrieval

State of the Art

Why document retrieval?

- Huge collections of documents available
- Need very precise queries to retrieve useful information
- Only a few documents can be processed by users
- Typical example: Web search engines

How is it solved today?

- Inverted indexes: Limit queries to words and phrases, and precompute all the answers
- Very successful on natural language text collections
- Behind every text search engine out there

Why general sequences?

- The term natural language is more restrictive than what can be imagined, excluding
  - Oriental languages like Chinese, Korean, Thai, Japanese (Kanji), Lao, Vietnamese...
  - Highly synthetic languages like Finnish, Hungarian, Japanese, German...
- Other areas where document retrieval is of interest:
  - Bioinformatics: bases, amino acids, genes, ...
  - Software repositories: modules, functions, ...
  - Chemoinformatics: formulas
  - Multimedia sequences: MIDI notes, ...

Why not using inverted indexes?

- In these cases, we have a general sequence of symbols
- The query may appear as any substring in those sequences
- The inverted index cannot precompute the answer to every query
- A different technology is needed

Why not using string matching indexes?

- Suffix trees and suffix arrays can spot the individual occurrences of any substring in a general sequence collection
- But these are too many compared to the answer we really want
- We need an extension of those structures that can directly answer document retrieval problems efficiently.

Which problems we want to solve?

- Document listing: List the documents where a pattern $P[1, m]$ appears.
- Document listing with frequency: As above, but also give the frequency of P in each answer document.
- Top-k document retrieval: Give the k documents where P appears most often (or using some other relevance measure)

Why is the talk about?

- I will focus on the algorithmic issues of the problem
- I will describe the most interesting solutions to the problems described
- Most are mapped to another family of problems called colored range retrieval: given an array of colors, find distinct colors in a range
- This has its own set of application areas:
  - Web mining: distinct visitors to a page, frequently visited pages, common queries, ...
  - Database tuning: frequent queries, frequently accessed tables, ...
  - Business intelligence: itemset mining, e.g. items frequently bought together
  - Social behavior: words used on tweets, sites visited, topics queried, “likes”, posters in blogs, ...
  - Bioinformatics again: pattern discovery, e.g. frequent patterns in sequences

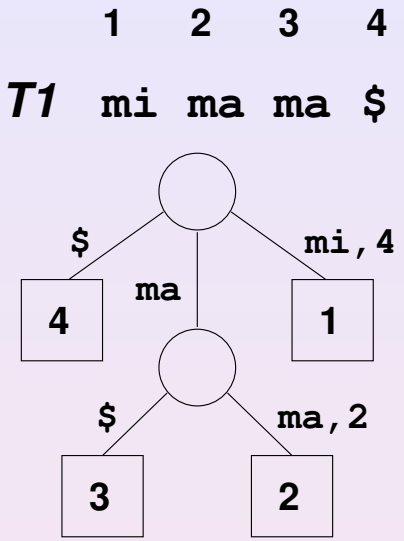
Why this Problem?

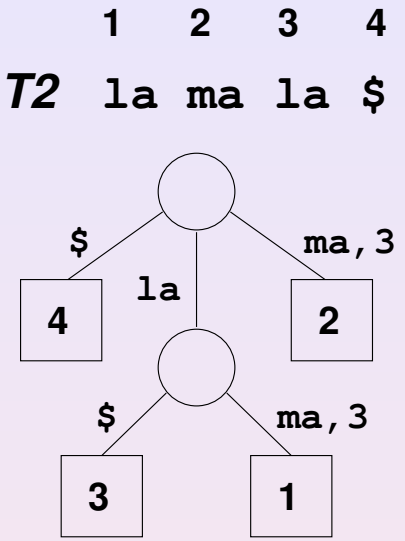
Document Listing

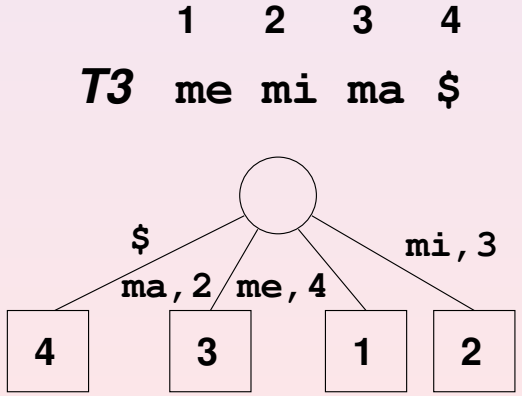
Top-k Retrieval

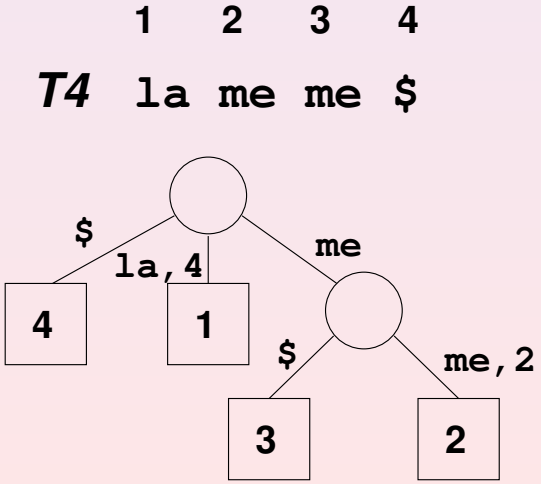
State of the Art

Suffix trees and suffix arrays



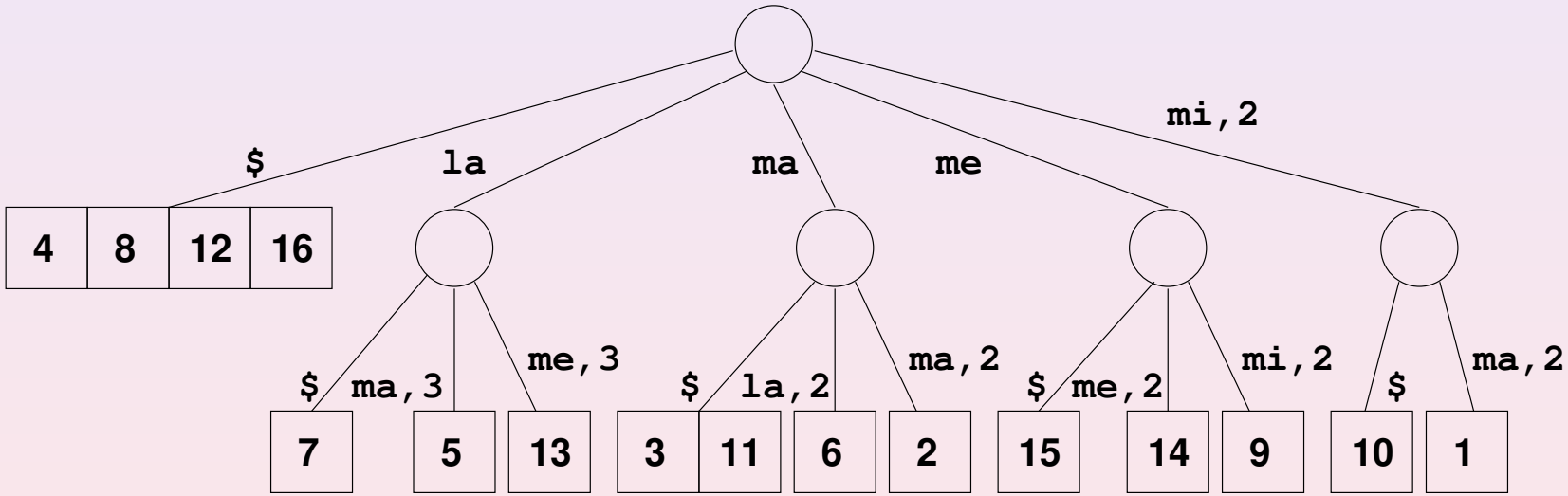






Suffix trees and suffix arrays

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	mi	ma	ma	$	la	ma	la	$	me	mi	ma	$	la	me	me	$



Suffix trees and suffix arrays

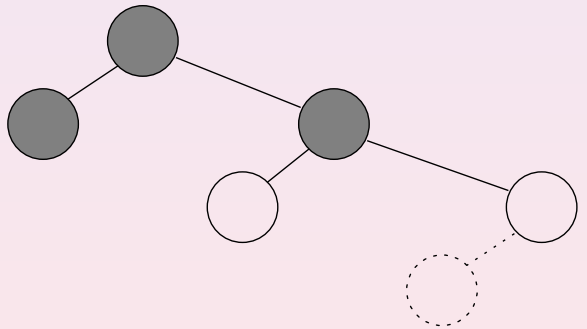
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	mi	ma	ma	$	la	ma	la	$	me	mi	ma	$	la	me	me	$

A																
	4	8	12	16	7	5	13	3	11	6	2	15	14	9	10	1

Muthukrishnan's algorithm: listing colors

Muthukrishnan's algorithm: listing colors

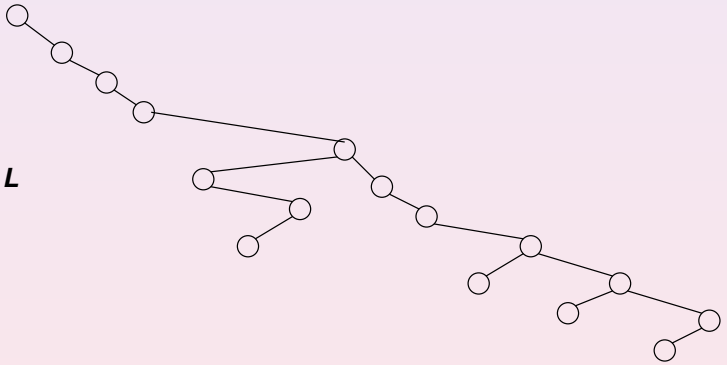
C	1	2	3	4	2	2	4	1	3	2	1	4	4	3	3	1
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
L	0	0	0	0	2	5	4	1	3	6	8	7	12	9	14	11



Sadakane's version: optimal space

T	mi	ma	ma	$	la	ma	la	$	me	mi	ma	$	la	me	me	$
B	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1

CSA	4	8	12	16	7	5	13	3	11	6	2	15	14	9	10	1



RMQ_L

V	0	0	0	0

L	0	0	0	0	2	5	4	1	3	6	8	7	12	9	14	11

Sadakane's version: term frequency

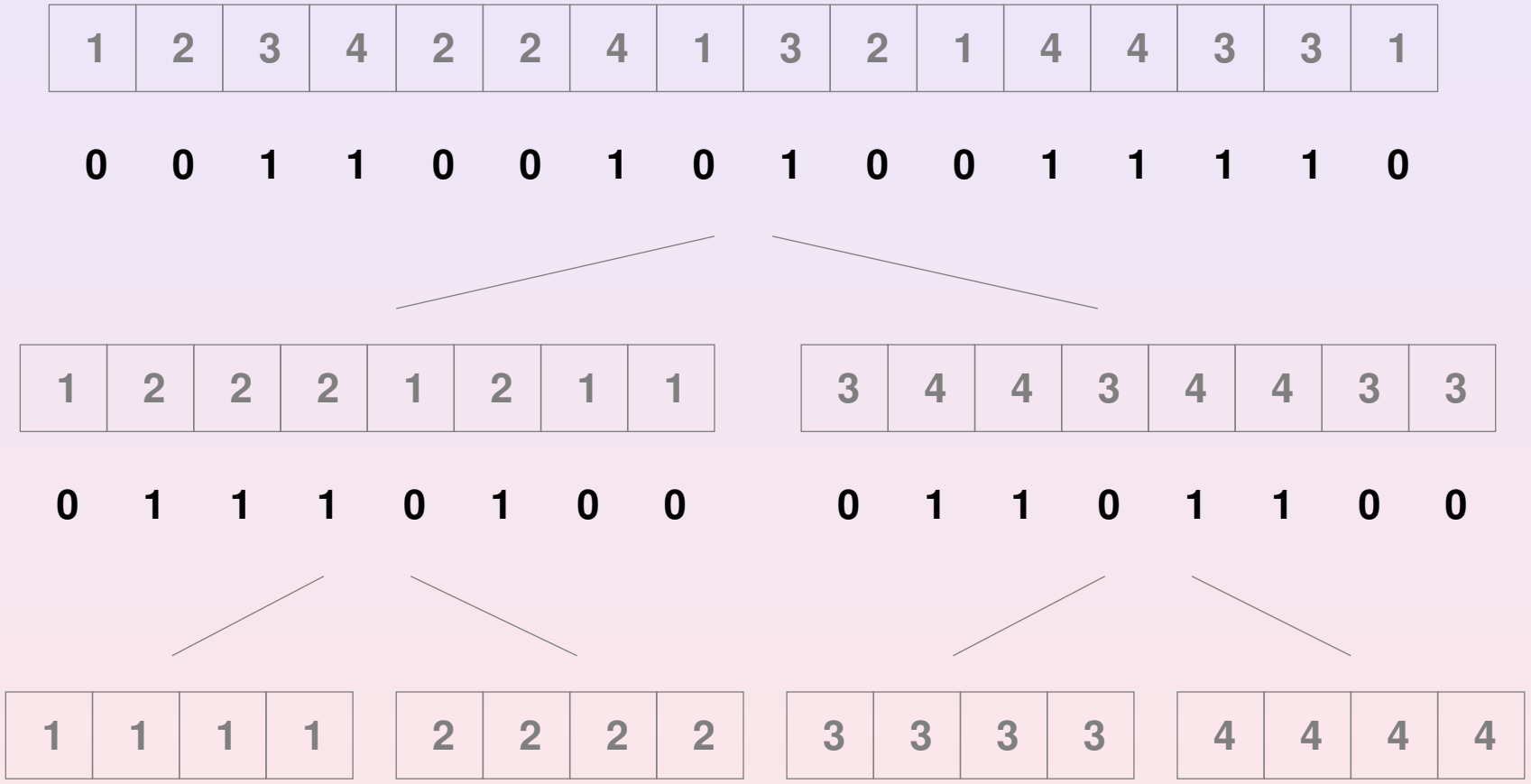
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
T	mi	ma	ma	$	la	ma	la	$	me	mi	ma	$	la	me	me	$
B	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1

A	4	8	12	16	7	5	13	3	11	6	2	15	14	9	10	1

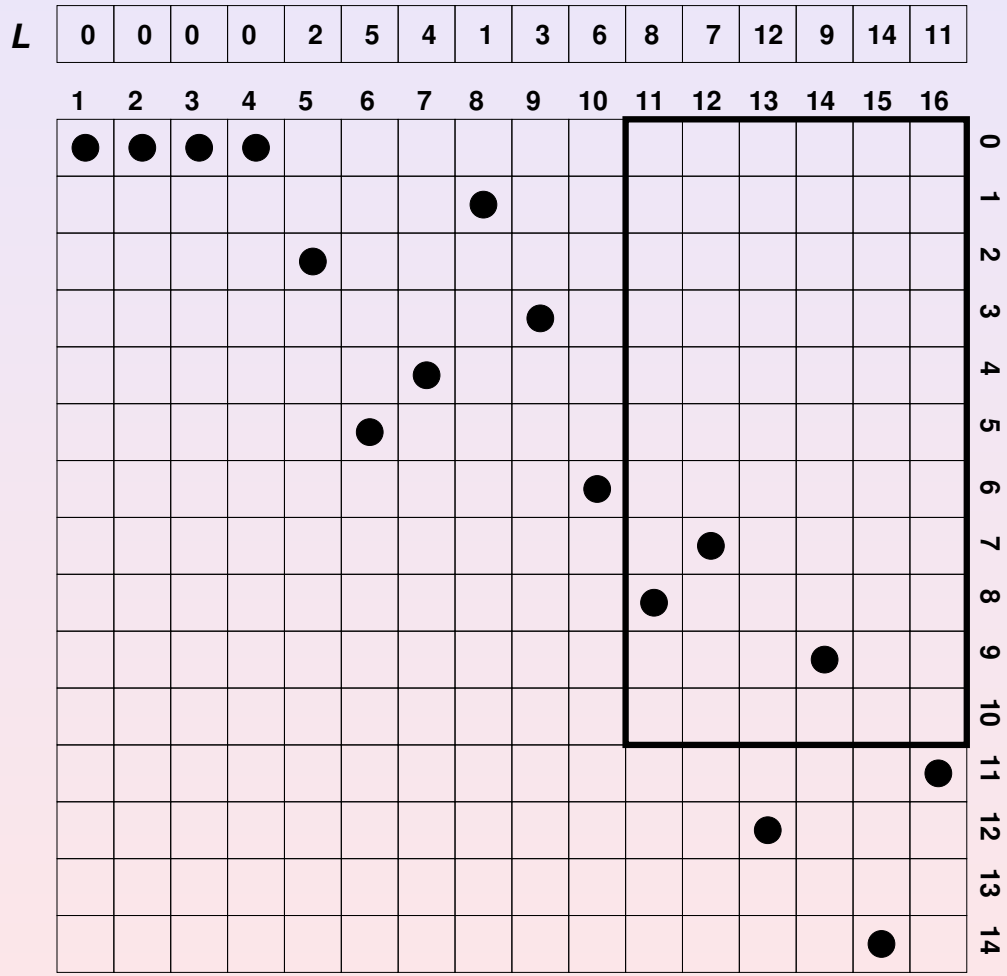
A_2	4	1	3	2

A_2^{-1}	2	4	3	1

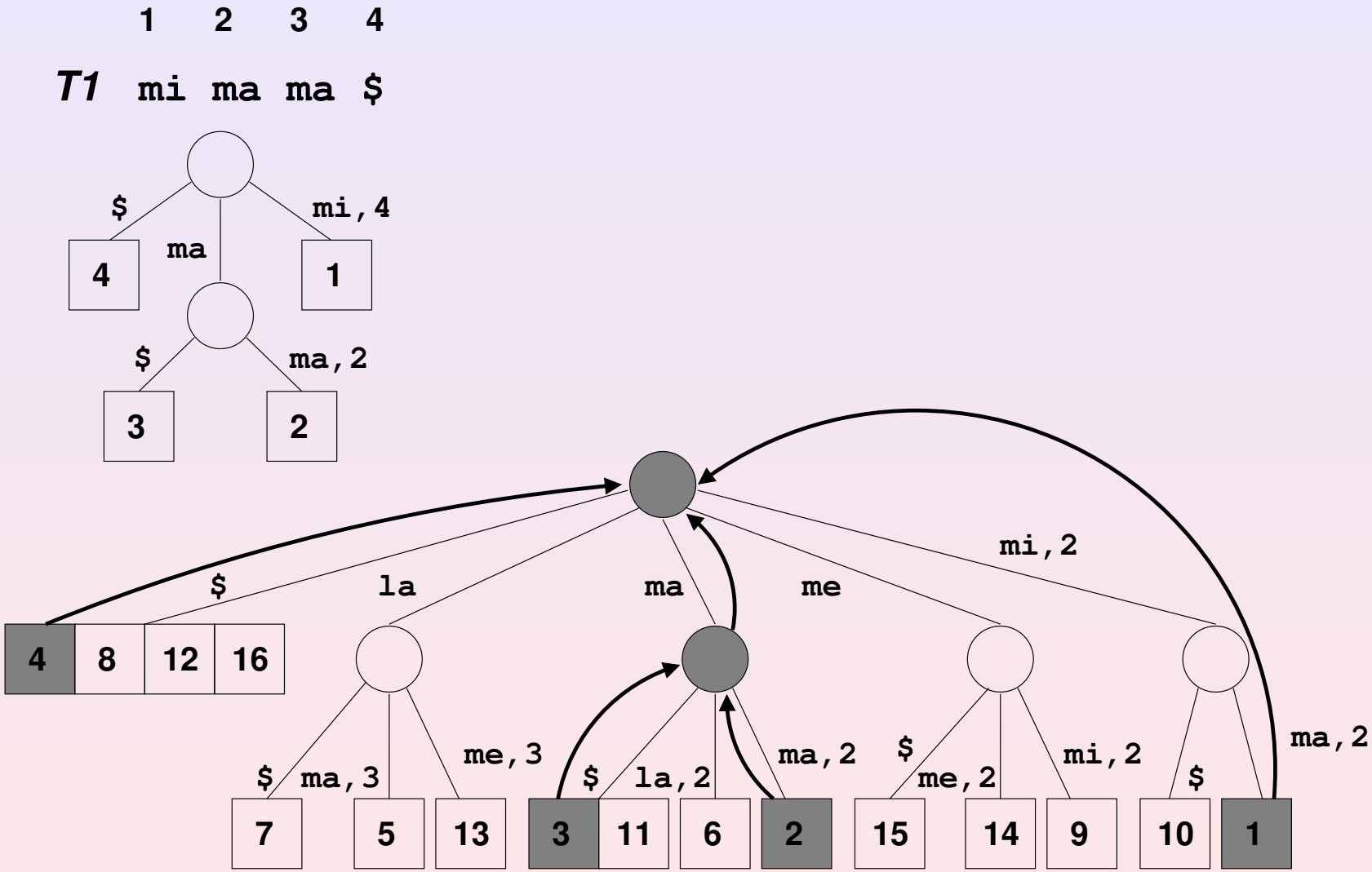
Wavelet trees: frequency and importance



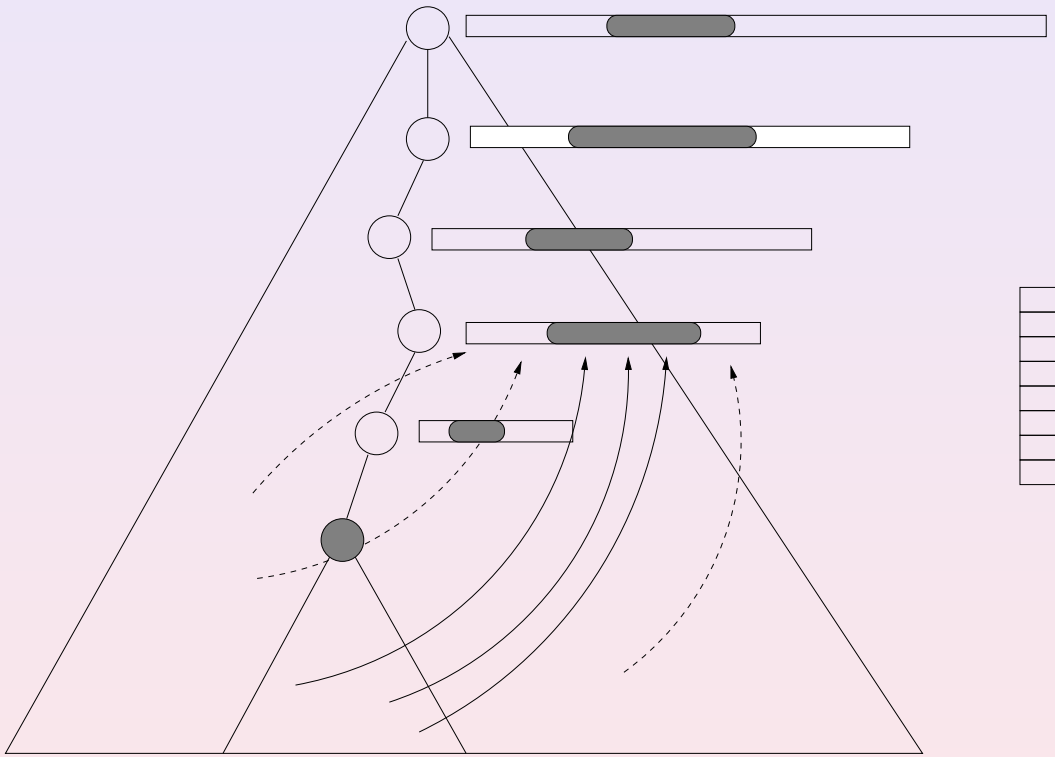
Document frequency: counting colors

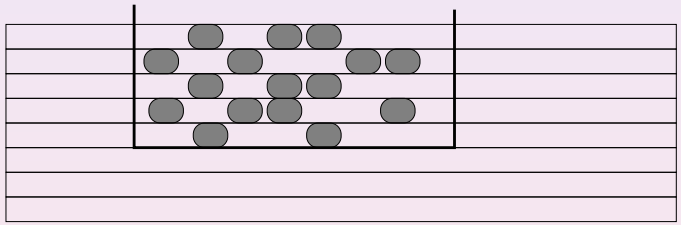


Hon, Shah, and Vitter's solution

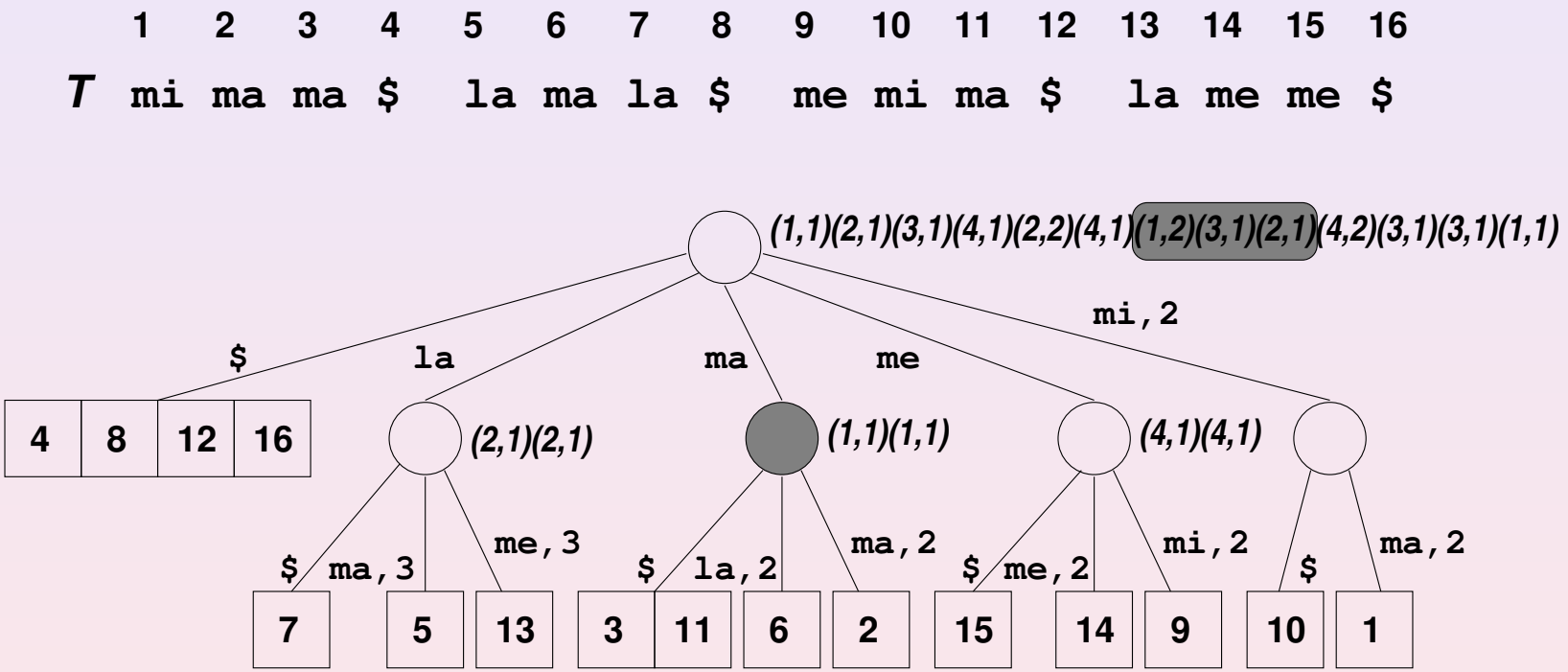


Hon, Shah, and Vitter's solution

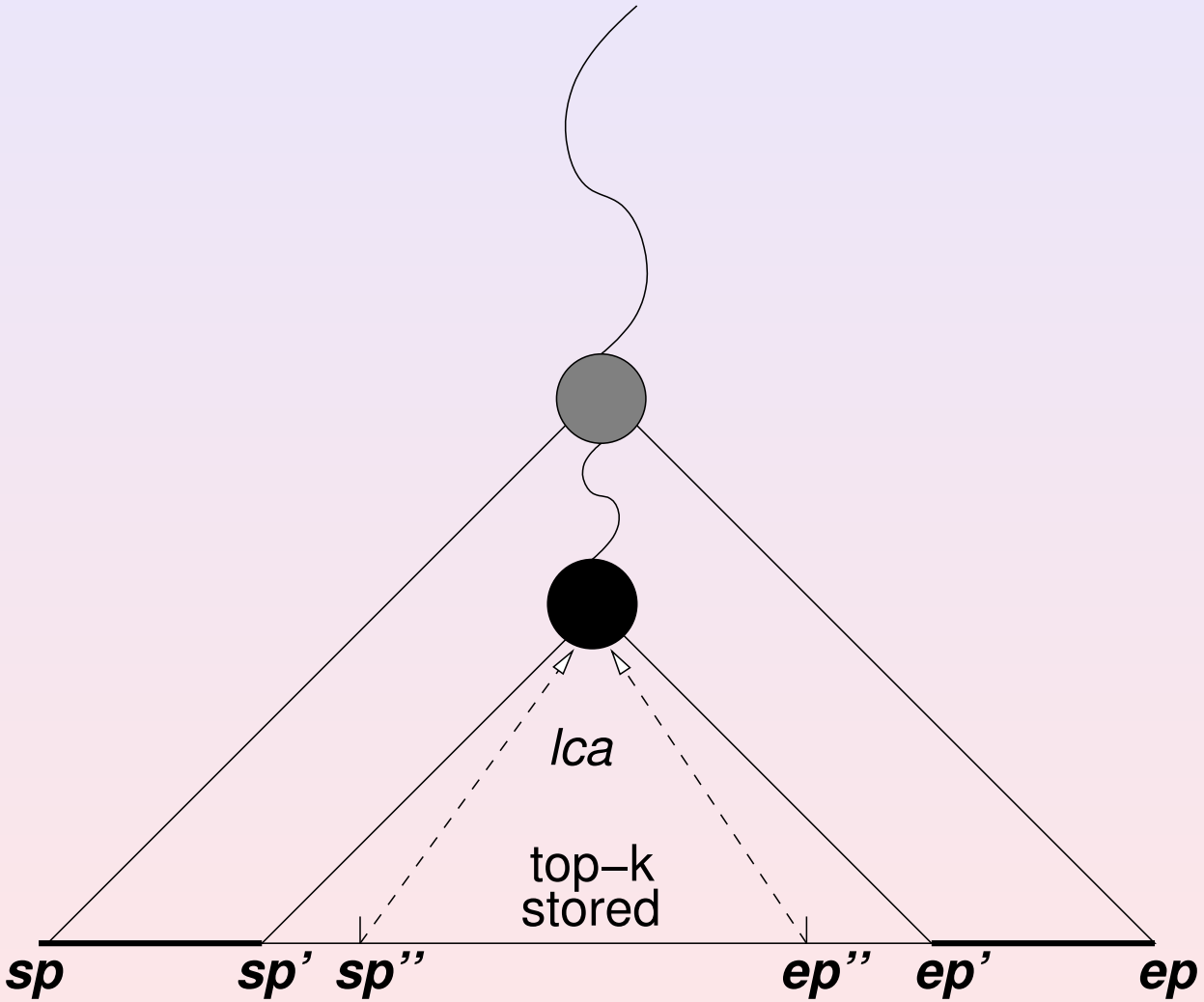




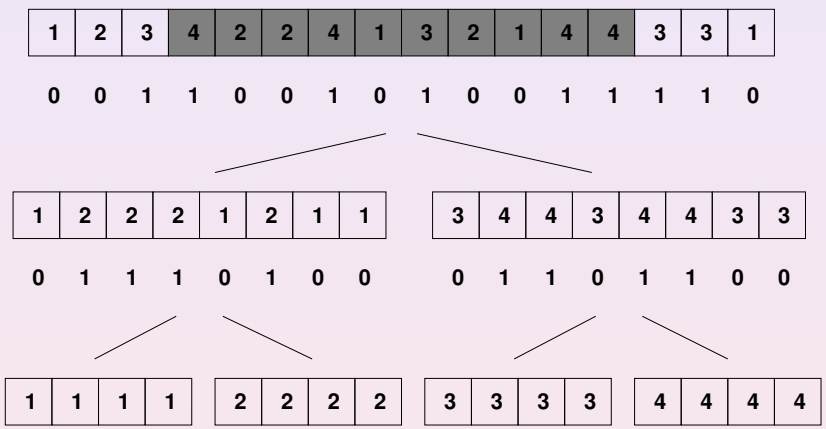
Hon, Shah, and Vitter's solution

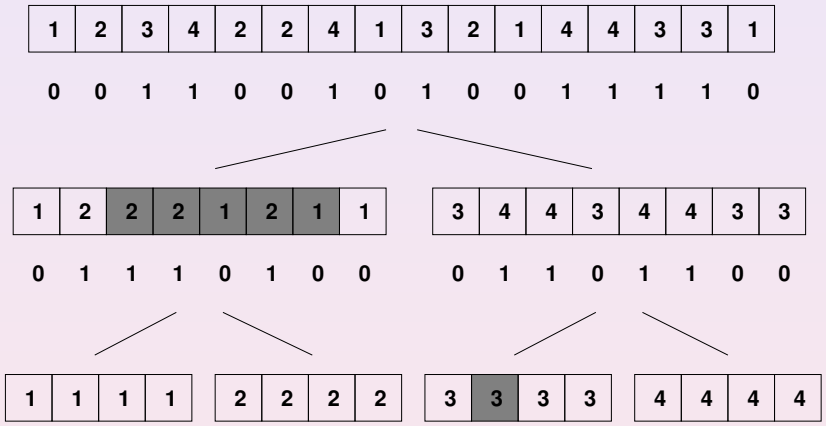


Top-k in reduced space



Top-k colors in general: hard





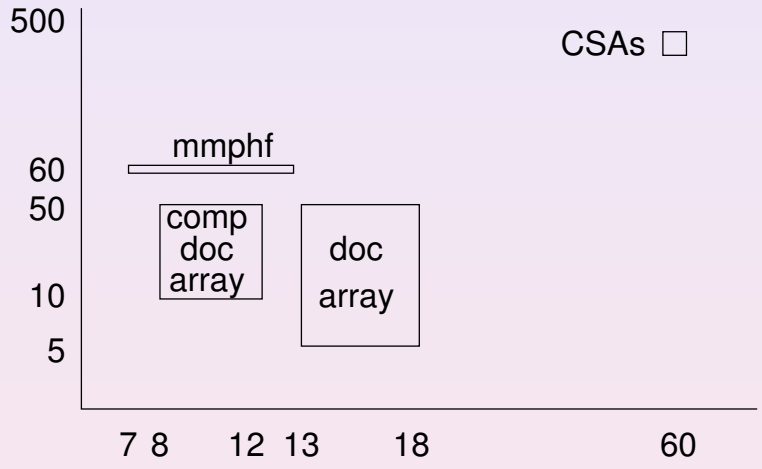
State of the art: Colors

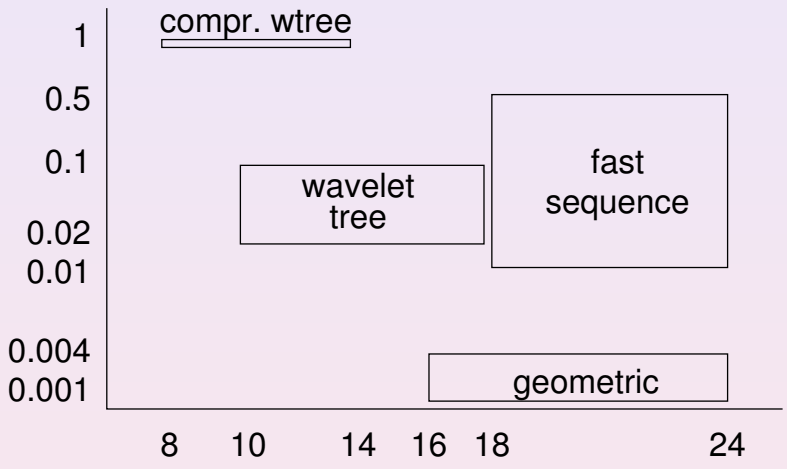
Problem		Time	Space
Color Listing	Sadakane 2007	Real	Data + $O(n)$
Color Listing with Freqs	Belazzougui et al. 2013	Optimal	Data + $O(n \lg \lg D + D \lg n)$
Color Counting	Gupta et al. 1995; Bose et al. 2009	$\lg n / \lg \lg n$	$n \lg n + o(n \lg n)$
	Gagie et al. 2013	$\lg(ep - sp + 1)$	$n \lg D + o(n \lg D) + O(n)$
Top-k Heaviest Colors	Karpinski and Nekrich 2011	Real	$O(n \lg D)$
	Gagie et al. 2012	$k \lg(D/k)$	$n \lg D + o(n \lg D)$ *
Top-k Colors ($(1+\epsilon)$-app.)	Gagie et al. 2013	$k \lg D \lg(1/\epsilon)$	$O((n/\epsilon) \lg D \lg n)$

State of the art: Documents

Problem		Time	Space
Document Listing	Sadakane 2007	1	$\|\mathrm{CSA}\| + n \lg D$
	Hon et al. 2009	$\lg^{1+\epsilon} n$	$\|\mathrm{CSA}\| + o(n)$
Document Listing with Frequencies	Belazzougui et al. 2013	1	$\|\mathrm{CSA}\| + n \lg D$
	Sadakane 2007	$\lg^{1+\epsilon} n$	$2\|\mathrm{CSA}\| + O(n)$
	Belazzougui et al. 2013	$\lg^{1+\epsilon} n$	$\|\mathrm{CSA}\| + O(n \lg \lg \lg D)$
	this survey	$\lg \mathrm{docc} \lg^{1+\epsilon} n$	$\|\mathrm{CSA}\| + o(n)$
Document Frequency	Sadakane 2007	1	$\|\mathrm{CSA}\| + O(n)$
Top-k Most Important Documents	Karpinski and Nekrich 2011	1	$\|\mathrm{CSA}\| + O(n \lg D)$
	Gagie et al. 2012	$\lg(D/k)$	$\|\mathrm{CSA}\| + n \lg D$
	Belazzougui et al. 2013	$\lg k \lg^{1+\epsilon} n$	$\|\mathrm{CSA}\| + o(n)$
Top-k Documents	N. and Nekrich 2012	1	$O(n \lg D + n \lg \sigma)$
	N. and Thankachan 2013	$\lg^* k$	$\|\mathrm{CSA}\| + n \lg D$
	Hon et al. 2013	$\lg k \lg^{1+\epsilon} n$	$2\|\mathrm{CSA}\| + o(n)$
	N. and Thankachan 2013	$\lg^2 k \lg^{1+\epsilon} n$	$\|\mathrm{CSA}\| + o(n)$

State of the art: Documents





State of the art: Future challenges

- Practice: reduce the space!
- Theory: find the optimal space/time tradeoffs
- Extensions: multi-pattern queries (difficult)
- Applications: back to natural language?

Spaces, Trees, and Colors: The Algorithmic Landscape of Document Retrieval on Sequences.
http://arxiv.org/abs/1304.6023