

Finite State Automata in *Lucene*

Dawid WEISS



Dawid Weiss

20+ years of coding

10 years assembly only

Academia & Research

PhD in Information Retrieval, PUT

Open source

Carrot², HPPC, Lucene, ...

Industry & Business

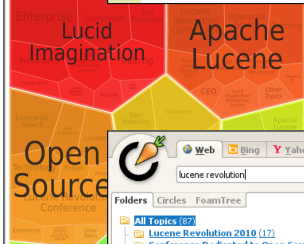
Carrot Search s.c.



Web Search

Concept-Driven Algorithm for Clustering Search Results

Stanisław Osński and Dawid Weiss, Poznań University of Technology



Top 87 results of

- 1 [Join the Lucene Revolution 2010](#)
Lucene Revolution 2010
<http://lucenero.com/>
- 2 [Agenda | www.lucene.org](#)
Lucene Revolution 2010
<http://www.lucene.org/>

Talk outline

State machines (automata)

FSA, DFA, FST and other XXXs.

Use cases in Lucene and Solr

Suggester. FuzzySearch. Index.

No API details

Still @experimental.

(Non)? Deterministic Finite State (Automata|Machines)

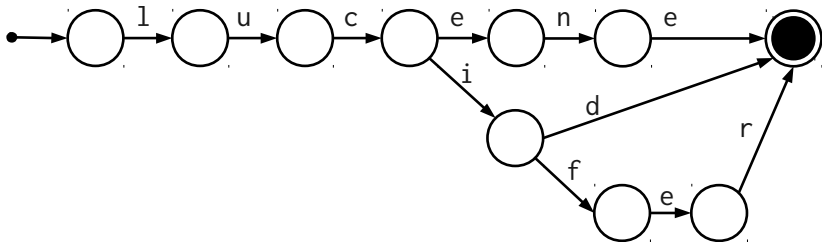
HashSet

hash	→ slot	→ value
0x29384d34		→ lucene
0xde3e3354		→ lucid
0x00000666		→ lucifer

HashSet

hash	→ slot	→ value
0x29384d34		→ lucene
0xde3e3354		→ lucid
0x00000666		→ lucifer

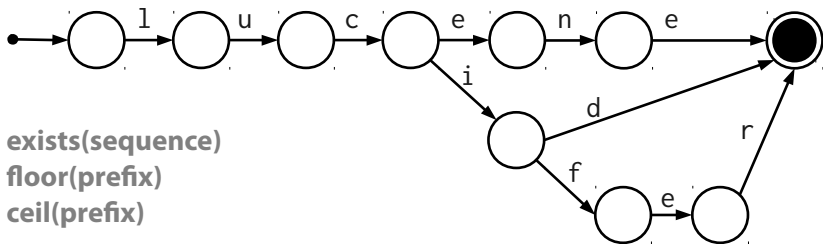
FSA (deterministic)

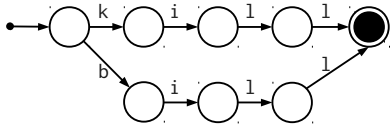


HashSet

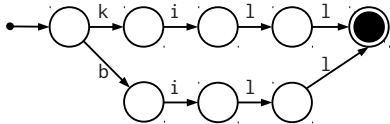
hash	→ slot	→ value
0x29384d34		→ lucene
0xde3e3354		→ lucid
0x00000666		→ lucifer

FSA (deterministic)

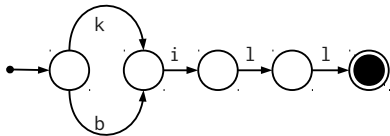




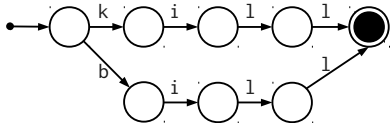
deterministic, non-minimal



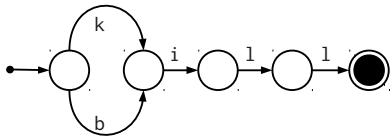
deterministic, non-minimal



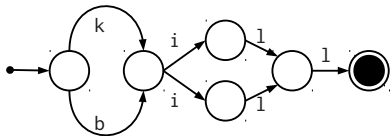
deterministic, minimal



deterministic, non-minimal



deterministic, minimal



**non-deterministic,
non-minimal**

(Sorted)Map

Lucene → 1
lucid → 2
lucifer → 666

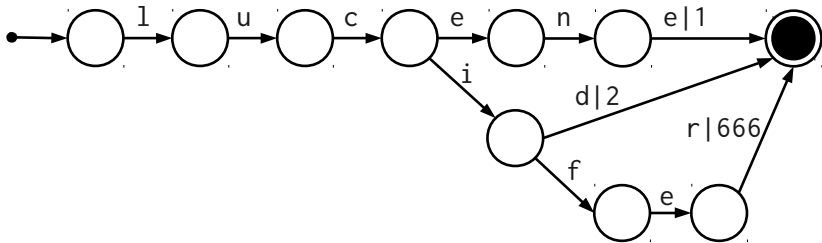
(Sorted)Map

Lucene → 1

Lucid → 2

Lucifer → 666

FST (transducer)



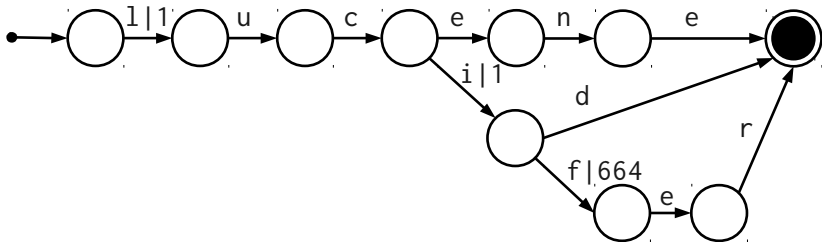
(Sorted)Map

Lucene → 1

Lucid → 2

Lucifer → 666

FST (transducer)



NFSAs and

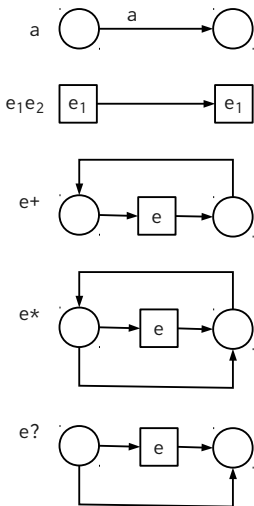
Regular expressions

Determinization

states explosion, not always possible

Backtracking

recursion explosion



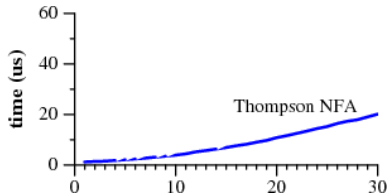
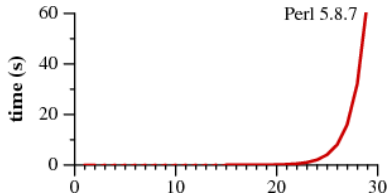
$a^n a^n$

$a^n a^n$

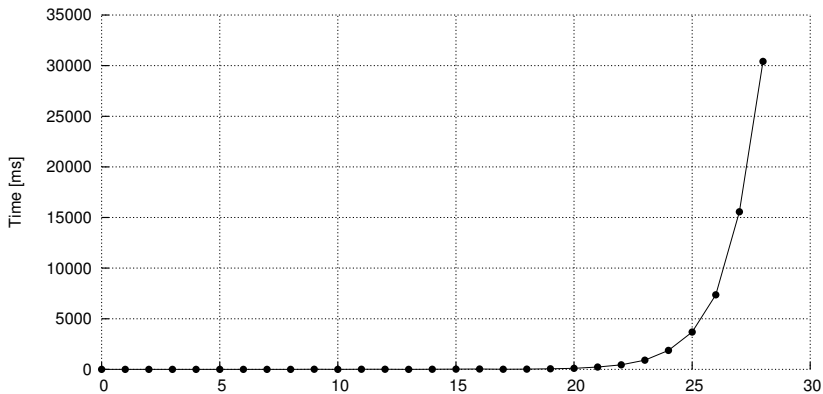
$n=3 \rightarrow a^3 a^3$

$a^n a^n$

$n=3 \rightarrow a^3 a^3$



Source: Russ Cox, Regular Expression Matching Can Be Simple And Fast (re2).



Time of matching a^n for pattern $a^n a^n$, depending on n . Java 1.6, modern hardware.

Linear-time, minimal, deterministic

FSA construction

Linear algorithm from sorted input

by Daciuk, Mihov, et al.

Active path

states that still can change

States dictionary

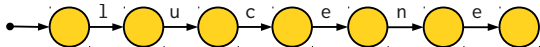
nodes that will never change

- 1) common AP prefix**
- 2) freeze the rest of AP**
- 3) add suffix → new AP**



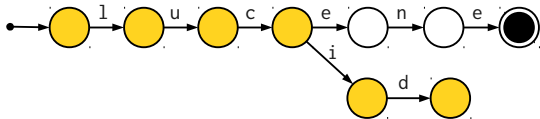
lucene

- 1) common AP prefix
- 2) freeze the rest of AP
- 3) add suffix → new AP

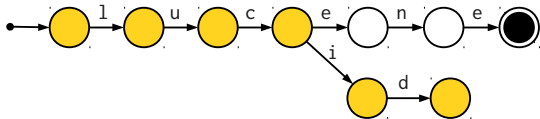


lucid

- 1) common AP prefix
- 2) freeze the rest of AP
- 3) add suffix → new AP

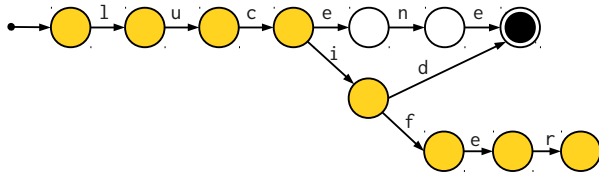


- 1) common AP prefix
- 2) freeze the rest of AP
- 3) add suffix → new AP

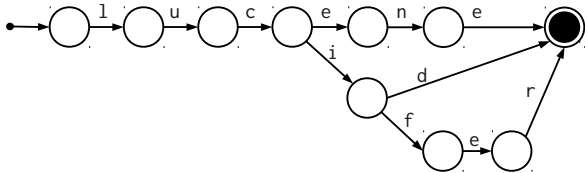


lucifer

- 1) common AP prefix
- 2) freeze the rest of AP
- 3) add suffix → new AP



- 1) common AP prefix
- 2) freeze the rest of AP
- 3) add suffix → new AP



FS(A|T)s in (Lucene|Solr)

Automata in

Lucene|Solr

org.apache.lucene.util.automaton.*

partial port of brics, FuzzyQuery, AutomatonTermsEnum

org.apache.lucene.util.automaton.fst.FST

FSA and FSTs from sorted data, suggester, indexes

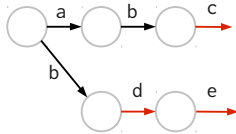
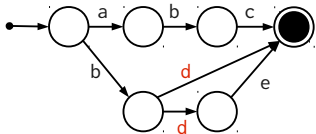
org.apache.lucene.util.automaton.fst.*

FSA representation

Arc-based, not state-based

Moore vs. Mealy. Compact vs. intuitive

Input: abc, bd, bde.



org.apache.lucene.util.automaton.fst.*

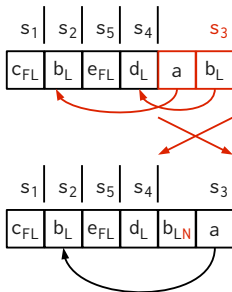
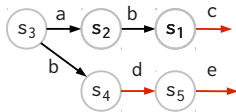
FSA representation

Arc-based, not state-based

Moore vs. Mealy. Compact vs. intuitive

Next-state chaining

requires unusual tricks during construction



org.apache.lucene.util.automaton.fst.*

FSA representation

Arc-based, not state-based

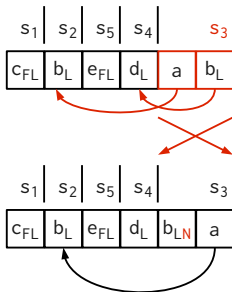
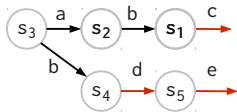
Moore vs. Mealy. Compact vs. intuitive

Next-state chaining

requires unusual tricks during construction

Everything in a byte[]

traversals-ready, memory-efficient



org.apache.lucene.util.automaton.fst.*

FSA representation

Arc-based, not state-based

Moore vs. Mealy. Compact vs. intuitive

Next-state chaining

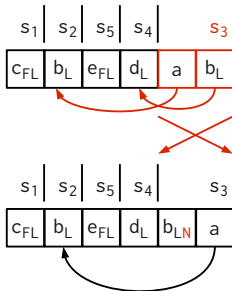
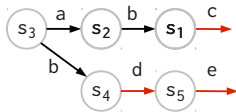
requires unusual tricks during construction

Everything in a byte[]

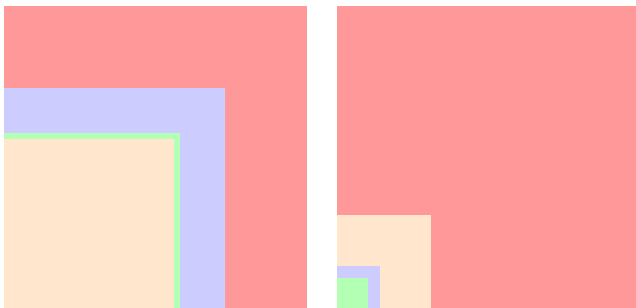
traversals-ready, memory-efficient

Dual transition storage format

lookup: bsearch or linear scan



Input	Input size		Compressed size (MB)		
	MB	Terms	Lucene	morf.	gzip
Wikipedia t.index	481	38 092 045	258	164	149
Polish infl.	162	3 672 200	3.1	1.7	15.4



Use Cases: Solr's Autocomplete

Solr's Suggesters

Design choices

sort order (alpha, score), prefix vs. spelling, boost exact matches?

Weights

term→weight, lookup(term, onlyMorePopular)

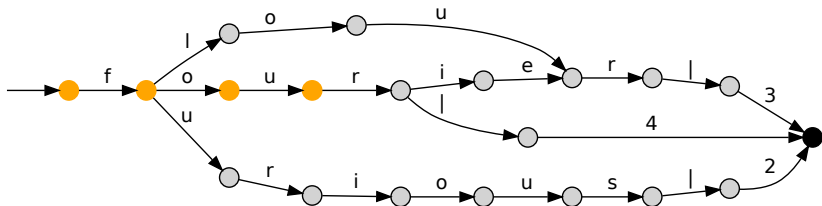
org.apache.solr.spelling.suggest.Lookup

JaspellLookup, TSTLookup, FSTLookup

flour|3
four|4
fourier|3
furious|2

f|our|3
fo|ur|4
fo|ur|ier|3
fo|ur|ious|2

→ fou*



Find prefix.

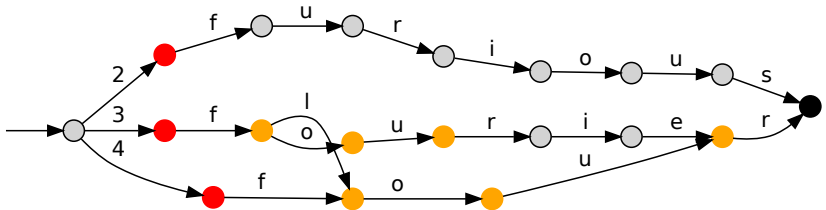
Depth-in traversal for completions.

PQ on score|alpha

2furious
3flour
3fourier
4four

2furious
3flour
3fourier
4four

→fou*



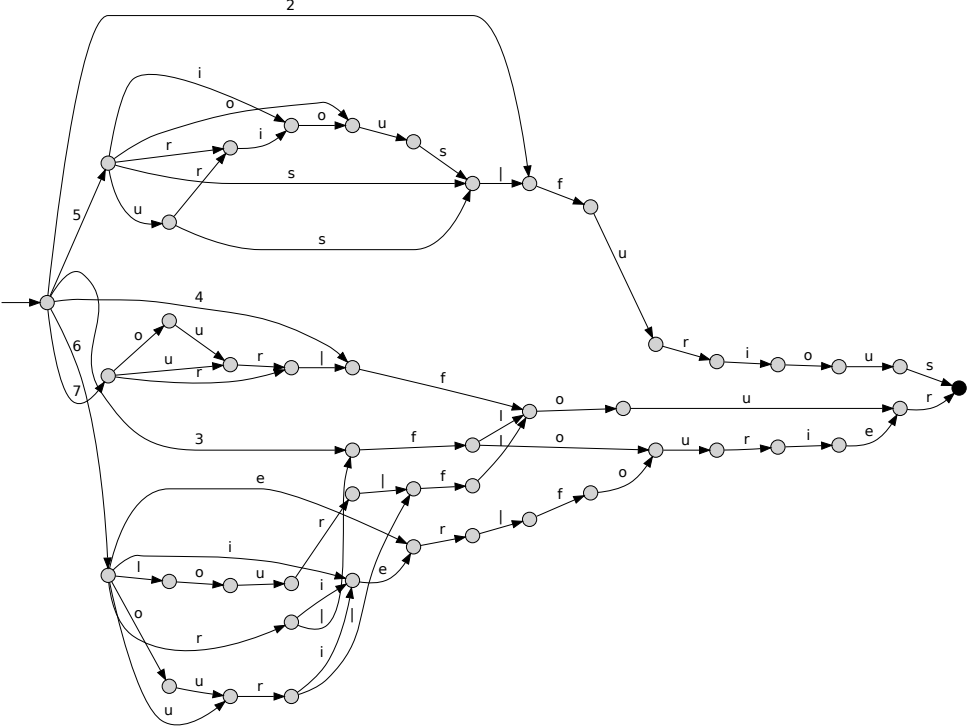
From score roots, until N collected.

Find prefix.

Depth-in traversal for completions, stop if N collected.

Find/boost exact match.

2furious
5urious|furious
5rious|furious
5ious|furious
5ous|furious
5us|furious
5s|furious
3flour
...



Constant time lookups!

Regardless of the terms dictionary size.

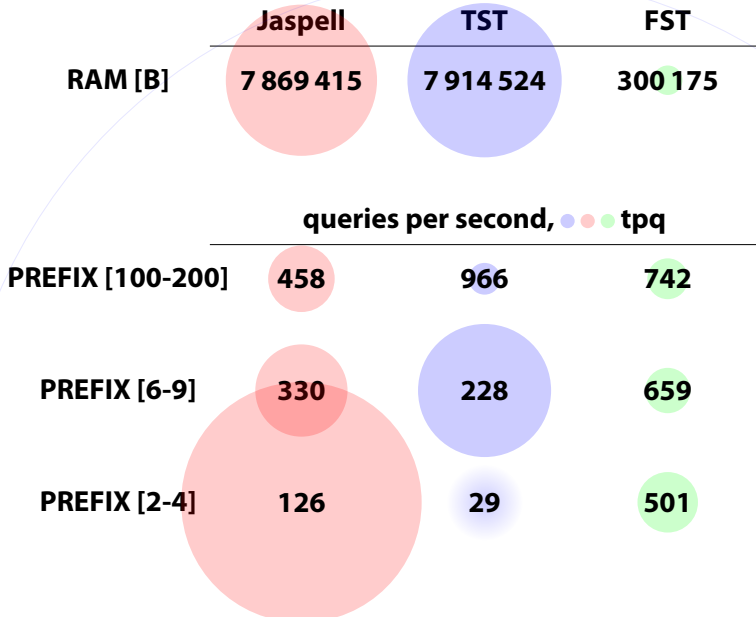
Regardless of prefix length.

Constant time lookups!

Regardless of the terms dictionary size.
Regardless of prefix length.

Exact matches only.
Static snapshot (not incremental).
Discretized weights.

Top50KWiki.utf8, 676 KB, 50 000 terms



Summary

Summary and Conclusions

Automata

compact, powerful, efficient data structure

Lucene/Solr benefits

behind the scenes, but spreading: index, queries, suggesters

API in Lucene

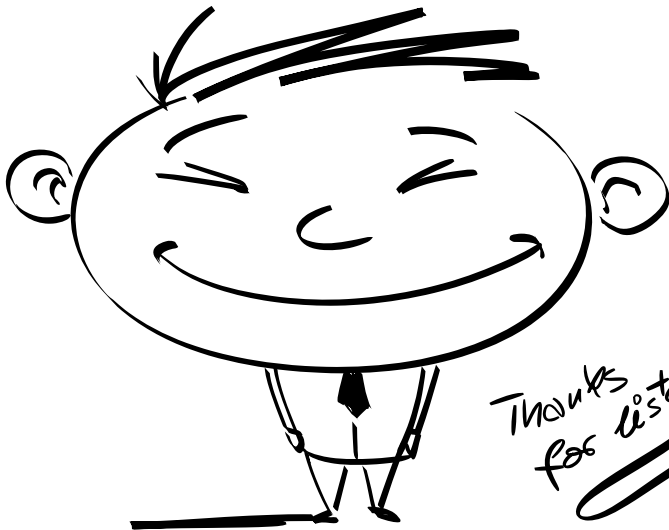
...is shaped right now, still @experimental

Acknowledgement

Michael McCandless

Robert Muir

committer: .+



Thanks for listening

dawid.weiss@carrotsearch.com