SIGIR 2016

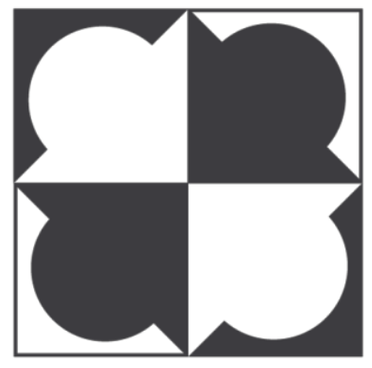
Richard Zanibbi, Kenny Davila, Andrew Kane, Frank W. Tompa

July 18, 2016

Multi-Stage Math Formula Search: Using Appearance-Based Similarity Metrics at Scale

R·I·T

UNIVERSITY OF WATERLOO



Mathematical Information Retrieval (MIR)

Many mathematical resources are available online, such as:

- Online databases of technical documents (papers, tutorials, instructional materials)
- **For non-experts:** Wikipedia, MathPlanet, Khan Academy
- **For experts:** On-line Encyclopedia of Integer Sequences (OEIS)

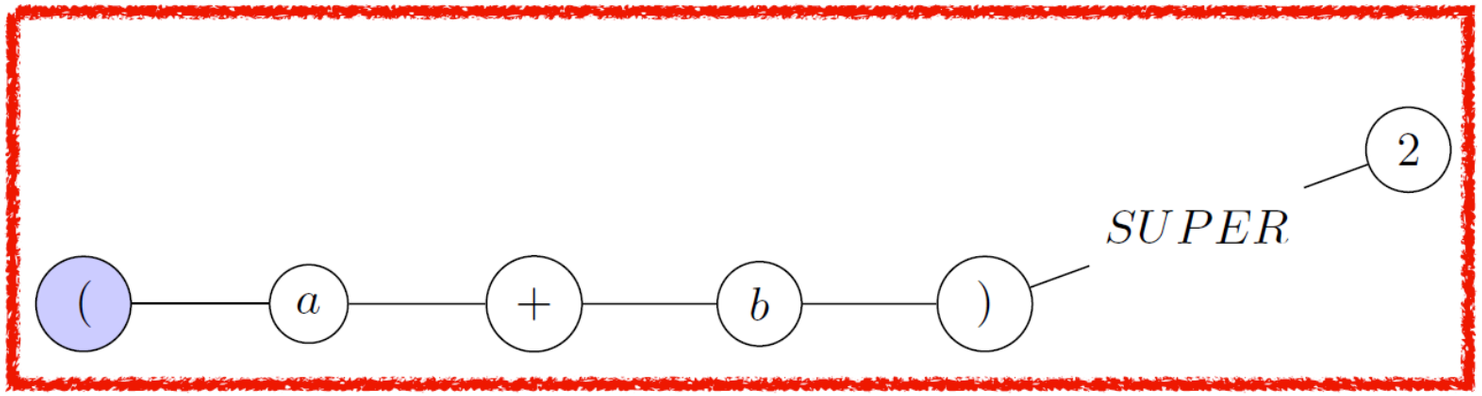
Unfortunately, formula search is not directly supported by major search engines; research is ongoing.

2014 Natural Research Council (USA) report on global digital math library initiative includes comments on existing search tool limitations.

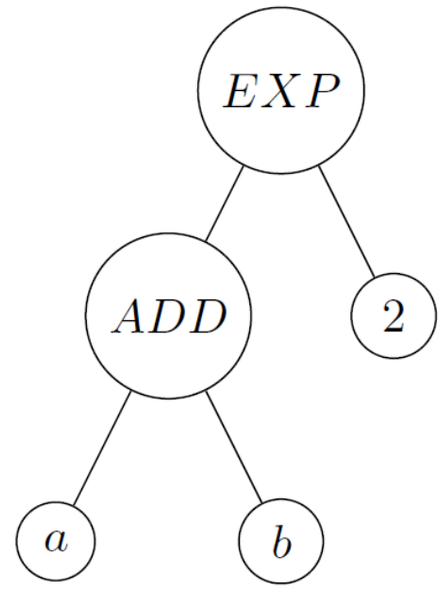
Surveys: [Zanibbi & Blostein, *IJDAR 2012*], [Guidi and Coen, *CICM 2015*]

See also papers from *NTCIR-10/-11/-12* Math Retrieval Tasks

Formula Encodings



a) Symbol layout tree (appearance)



b) Operator tree (semantics)

Translating between SLTs and OTs is heuristic,
due to dialectic usage of math notation

In our work, we consider appearance encodings
(LaTeX and Presentation MathML - commonly available)

Related Work

Approaches to Query-By-Expression may be categorized by the retrieval primitives used.

Text-based: tokens (linearized tree)

DLMF [Miller and Youssef, *Ann. Math. Artif. Intell. 38(1), 2003*]

Tree-based: complete sub-expressions

Substitution Index Trees [Kohlhase & Sucan, *AISC 2006*]

Tree Edit Distances [Kamali & Tompa, *CICM 2010, SIGIR 2013*]

Subexpression Hashing [Ohashi, Kristianto, Topic and Aizawa, *IEICE TOI&S, 2016*]

WikiMirs [X. Lin et al., *SIGIR 2014 ;* L. Gao et al., *NTCIR-12, 2016*]

Spectral: small structural units (e.g. paths)

Operator-Argument Triples [Nguyen, Hui and Chang, *Expert Sys. Appl., 2012*]

Bags of Paths [Hiroya and Saito, *NTCIR-10, 2013*]

Bags of Symbol Pairs (**Tangent**) [Stalnaker *MSc thesis, 2013;* Pattaniyil and Zanibbi, *NTCIR-11, 2014*]

Limitations of Tangent-2:
Spectral, Symbol Pair-Based Retrieval

---

1. Matched pairs may be **scattered** throughout an expression

2. **Exact matches have low similarity scores** within large expressions

3. **Pairs of wildcards are not indexed**, to avoid matching *all pairs* at ***δx, δy***

4. **Wildcards match only individual symbols** (e.g., 2^{-a} *partial* match for 2*)

5. **No unification** for symbols of the same type (e.g., x^2 and a^2)

6. Subexpression **groupings inconsistent** for matrices, vectors, parentheses, roots, etc., limiting recall

7. Retrieval is **slow,** even with parallelized retrieval (~ 5 secs/query); **all pairs of symbols along paths are indexed + symbols at end of writing lines (EOL).**

We address these limitations using a two-stage retrieval architecture.

tangent

Graphs

$O(*\log *)$

Search

Returned 41 matches (100 formulae, 153 docs)

Lookup 20.344 ms, Re-ranking 47.042 ms

Found 293777 tuple postings, 30342 formulae, 9916 documents

[formulas] [documents] [documents-by-formula]

1.0000
0.0000
3.0000

$O(E \log V)$ $O(M \log N)$ $O(d \log n)$

$O(k \log n)$ $O(\log \log N)$ $O(\log \log n)$

$O(\log \log y)$ $O(m \log n)$ $O(n \log h)$

$O(n \log k)$ $O(n \log m)$

1.0000
0.0000
2.0000

$\mathcal{O}(n \log \sigma)$

1.0000
0.0000
2.0000

$O(lgn)$ $O(MST)$ $O(KNM)$

Green: identical

Orange: unified

Red: wildcard

Black: unmatched

tangent

Graphs

$O(*\log *)$

Search

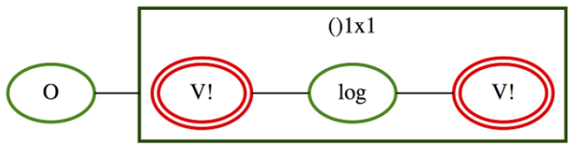
Returned 41 matches (100 formulae, 153 docs)

Lookup 20.344 ms, Re-ranking 47.042 ms

Found 293777 tuple postings, 30342 formulae, 9916 documents

[formulas] [documents] [documents-by-formula]

1.0000
0.0000
3.0000



$O(E \log V)$

$O(M \log N)$

$O(d \log n)$

$O(k \log n)$

$O(\log \log N)$

$O(\log \log n)$

$O(\log \log y)$

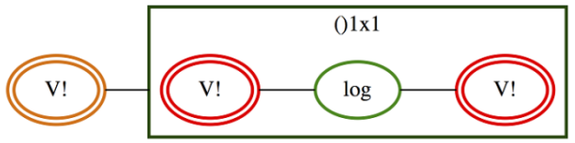
$O(m \log n)$

$O(n \log h)$

$O(n \log k)$

$O(n \log m)$

1.0000
0.0000
2.0000



$\mathcal{O}(n \log \sigma)$

Green: identical

Orange: unified

Red: wildcard

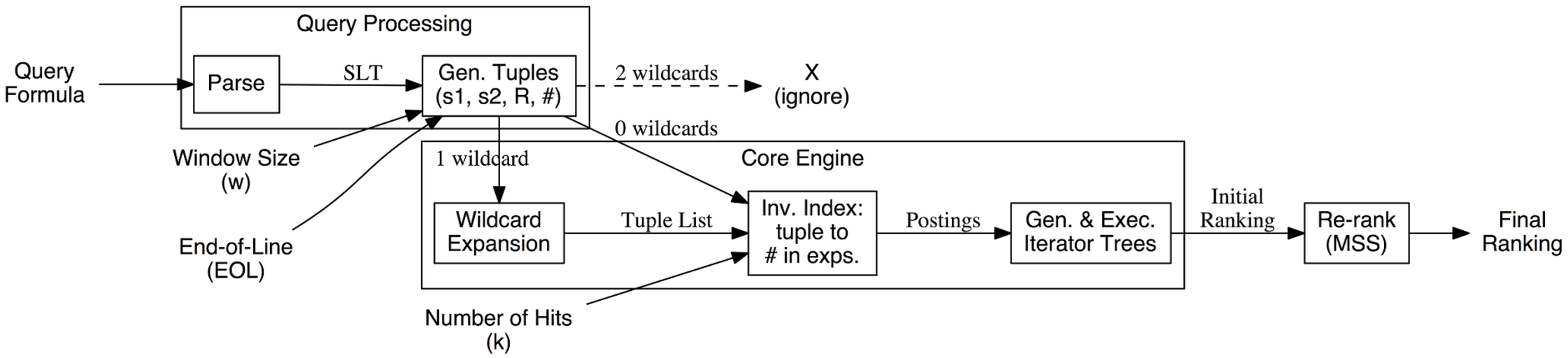
Black: unmatched

Tangent-3 Formula Search Engine

Tangent-3 Formula Retrieval Model

Steps

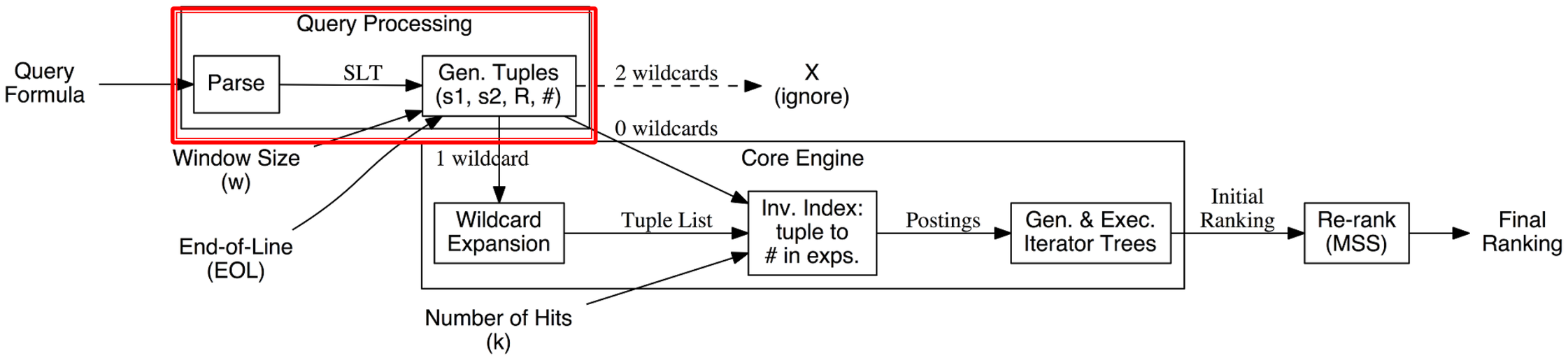
1. **Query Processing:** symbol pair generation
2. **Core Engine:** symbol pair retrieval
3. **Re-rank:** Max. Subtree Similarity (MSS)



Tangent-3 Formula Retrieval Model

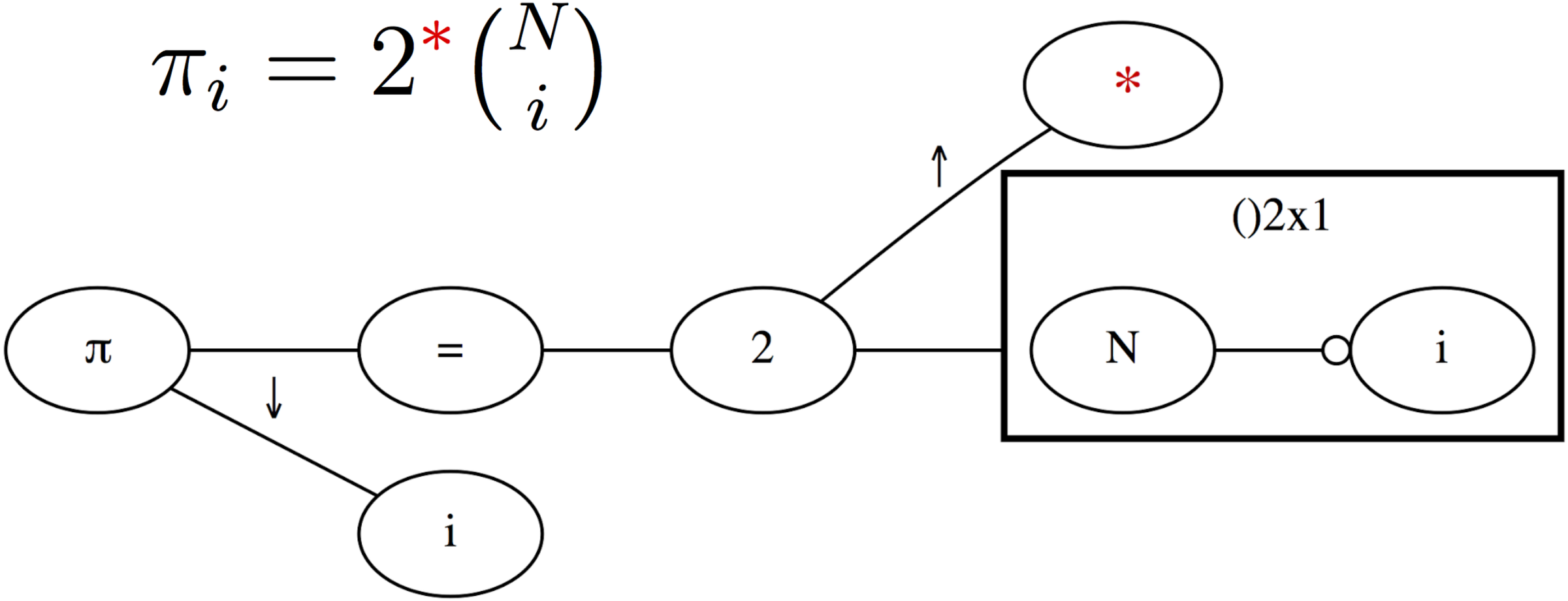
Steps

1. **Query Processing:** symbol pair generation
2. **Core Engine:** symbol pair retrieval
3. **Re-rank:** Max. Subtree Similarity (MSS)



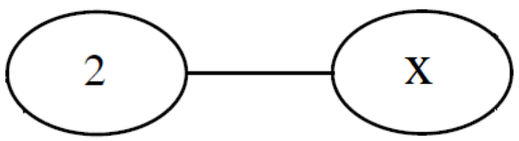
Symbol Layout Tree (SLT)

$$\pi_i = 2^{*}\binom{N}{i}$$



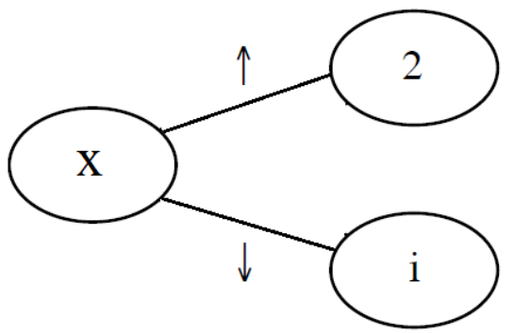
Symbol Layout Tree - Relationships

$2x$



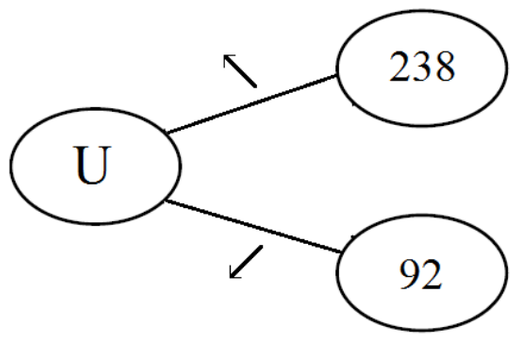
Next

X_i^2



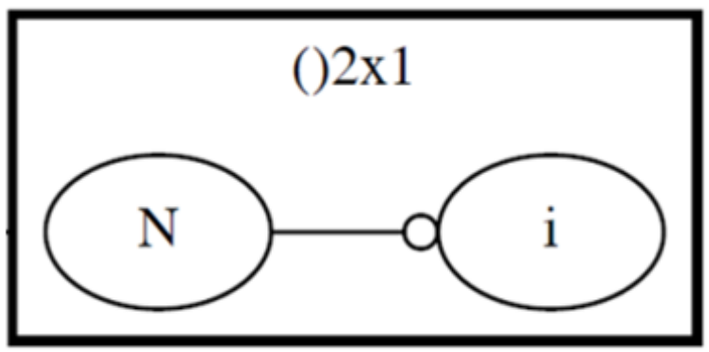
Above / Below

$^{238}_{92}\mathrm{U}$



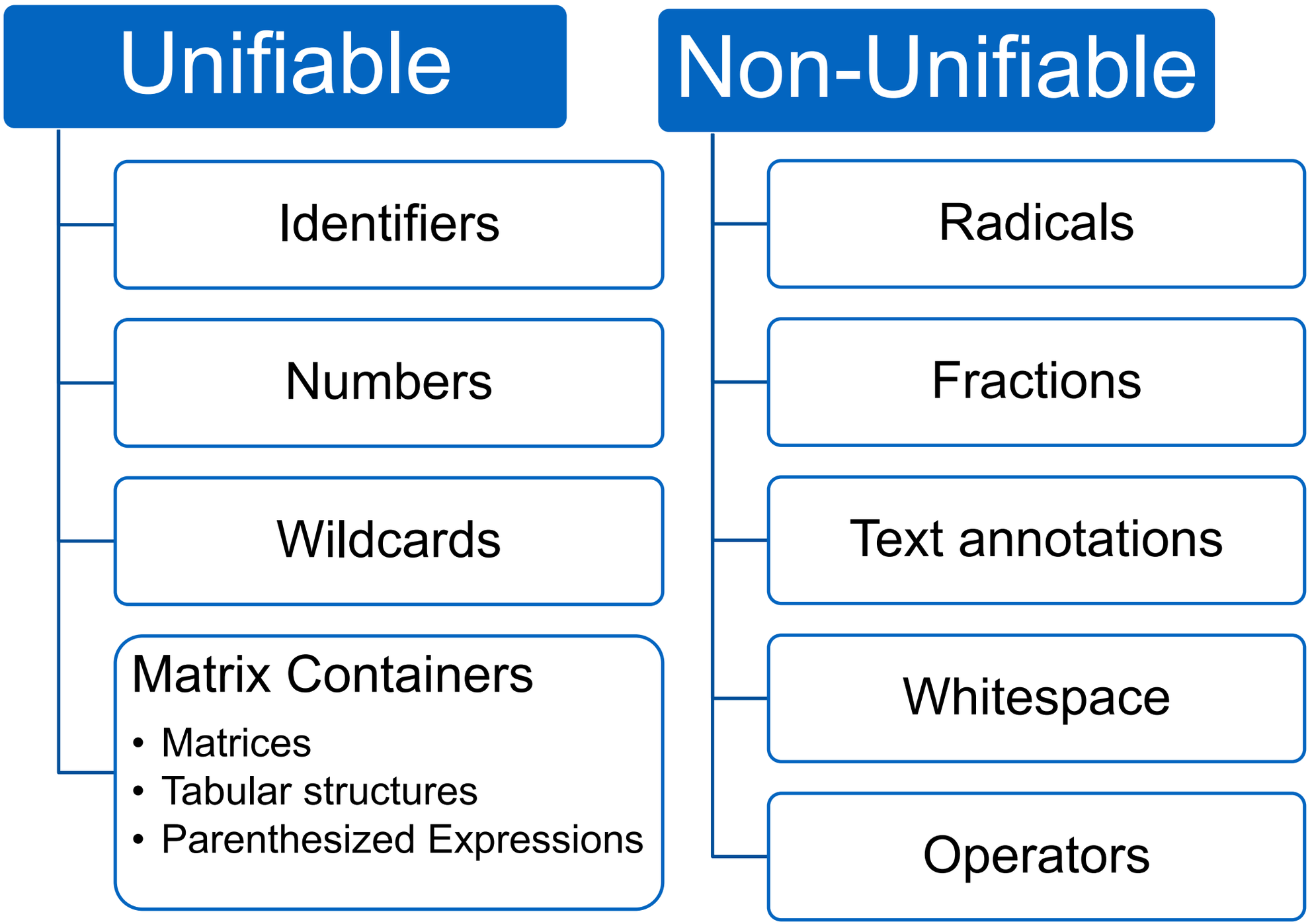
Pre-above / Pre-below

$\binom{N}{i}$



Within / Element

Symbol Layout Tree – Node Types



SLT Symbol Pair Generation

Tuples Generated

$(\boldsymbol{s_1}, \boldsymbol{s_2}, \boldsymbol{R}, \#)$

$\boldsymbol{s_1}$ - Ancestor Symbol

$\boldsymbol{s_2}$ - Descendent Symbol

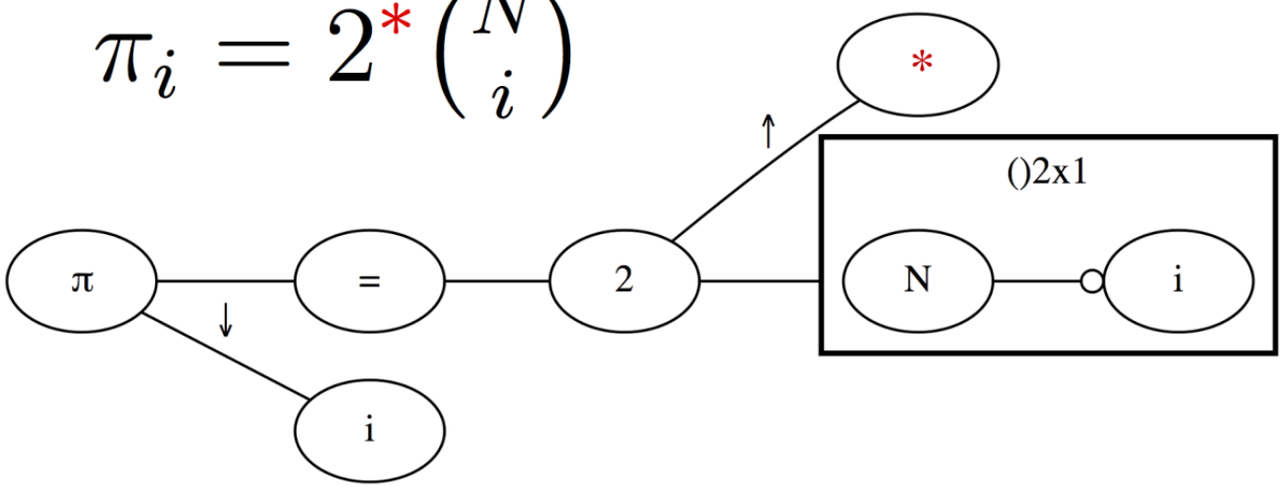
$\boldsymbol{R}$ - Edge Label Sequence between $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$

\# - Count

Parameters

- Window Size (w)
- End of Line

$$\pi_i = 2^* \binom{N}{i}$$

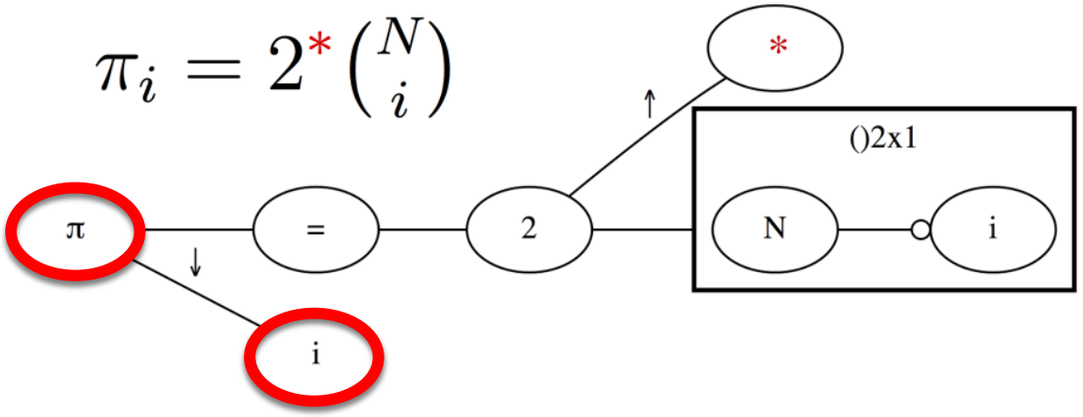


SLT Symbol Pair Generation

SYM-1	SYM-2	PATH	COUNT
V!π	**V!**i	$\downarrow$	1

Window size = 1

$$\pi_i = 2^* \binom{N}{i}$$

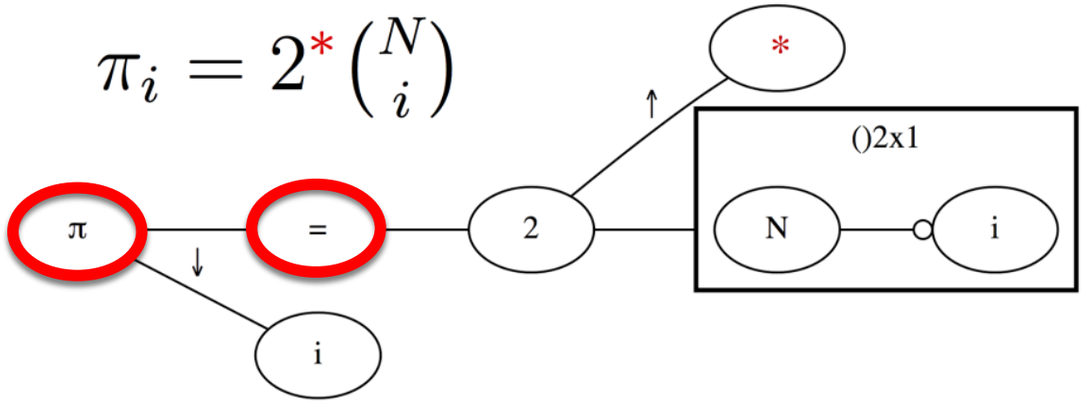


SLT Symbol Pair Generation

Window size = 1

SYM-1	SYM-2	PATH	COUNT
V!π	**V!**i	$\downarrow$	1
V!π	$=$	$\rightarrow$	1

$$\pi_i = 2^{*}\binom{N}{i}$$

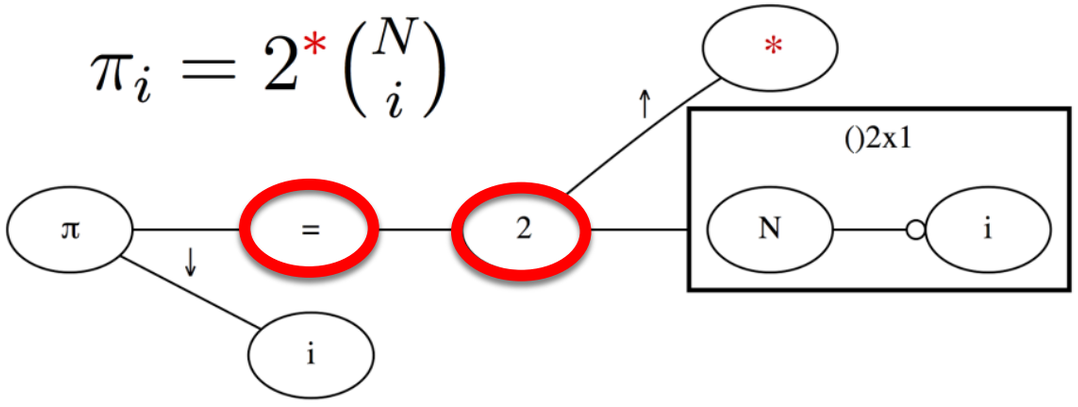


SLT Symbol Pair Generation

Window size = 1

Sym-1	Sym-2	Path	Count
$\mathbf{V!}\pi$	$\mathbf{V!}$i	$\downarrow$	1
$\mathbf{V!}\pi$	$=$	$\rightarrow$	1
$=$	$\mathbf{N!}2$	$\rightarrow$	1

$$\pi_i = 2^* \binom{N}{i}$$

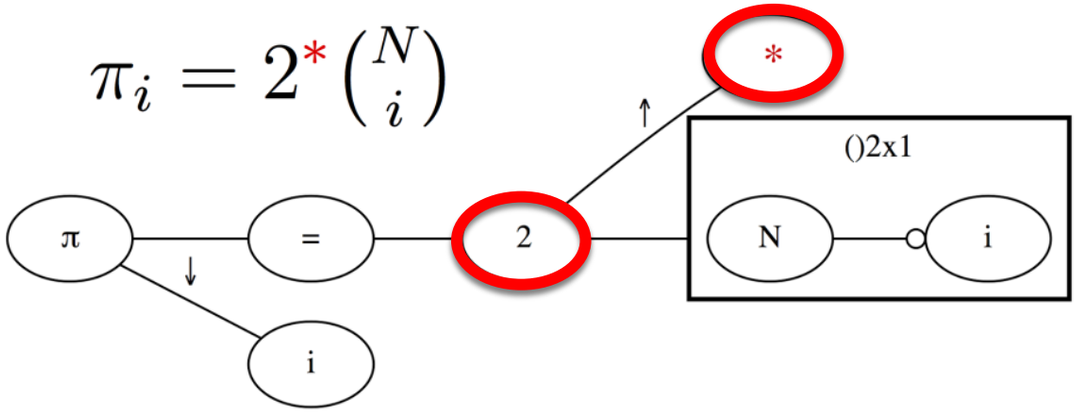


SLT Symbol Pair Generation

Window size = 1

Sym-1	Sym-2	Path	Count
V!π	**V!**i	$\downarrow$	1
V!π	$=$	$\rightarrow$	1
$=$	**N!**2	$\rightarrow$	1
N!2	*	$\uparrow$	1

$$\pi_i = 2^{*}\binom{N}{i}$$

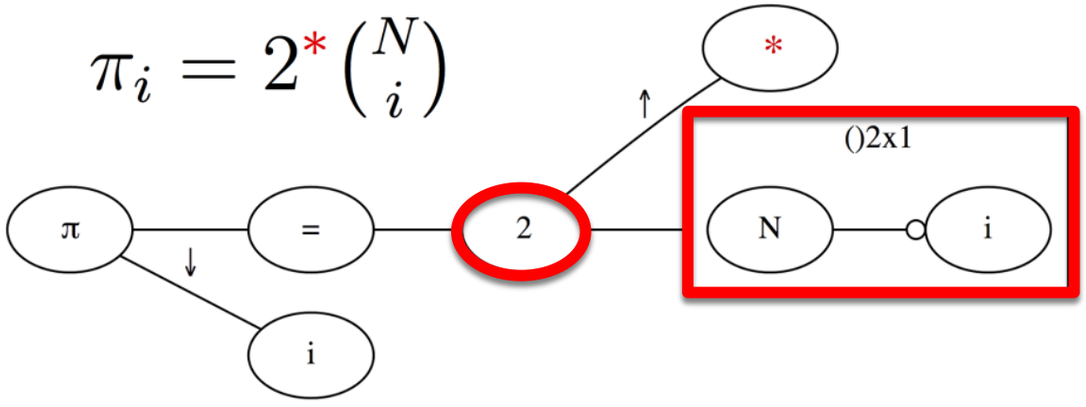


SLT Symbol Pair Generation

Window size = 1

Sym-1	Sym-2	Path	Count
$\mathbf{V!}\pi$	$\mathbf{V!}$i	$\downarrow$	1
$\mathbf{V!}\pi$	$=$	$\rightarrow$	1
$=$	$\mathbf{N!}2$	$\rightarrow$	1
$\mathbf{N!}2$	$*$	$\uparrow$	1
$\mathbf{N!}2$	$\mathbf{M!}()2\text{x}1$	$\rightarrow$	1

$$\pi_i = 2^* \binom{N}{i}$$

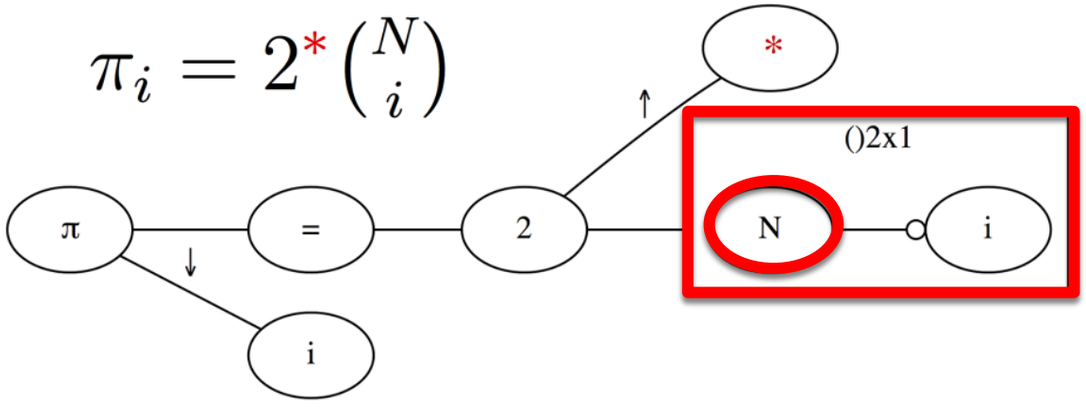


SLT Symbol Pair Generation

Window size = 1

Sym-1	Sym-2	Path	Count
V!π	**V!**i	$\downarrow$	1
V!π	$=$	$\rightarrow$	1
$=$	**N!**2	$\rightarrow$	1
N!2	*	$\uparrow$	1
N!2	**M!**()2x1	$\rightarrow$	1
M!()2x1	**V!**N	$\square$	1

$$\pi_i = 2^{*}\binom{N}{i}$$

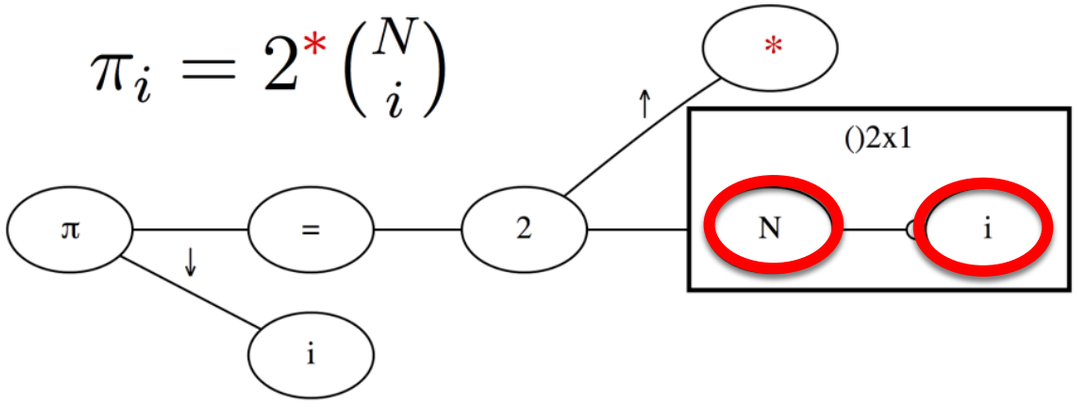


SLT Symbol Pair Generation

Window size = 1

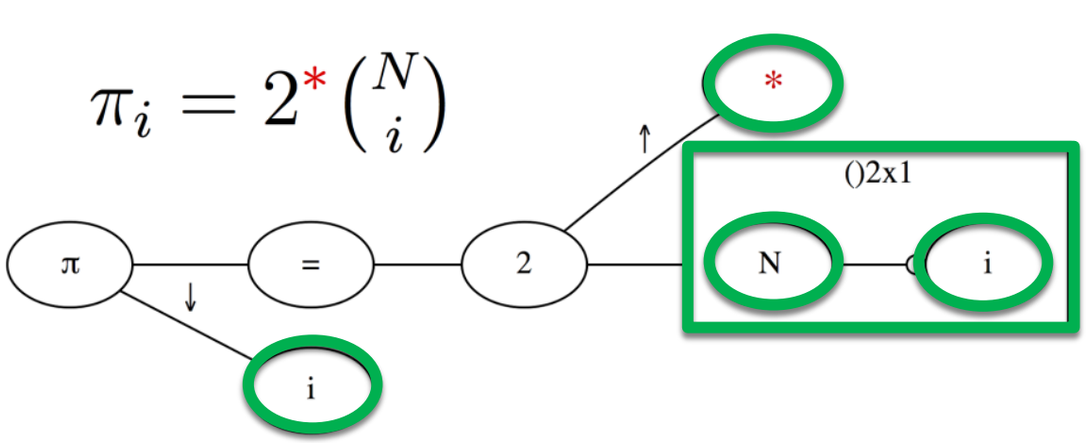
Sym-1	Sym-2	Path	Count
$\mathbf{V!}\pi$	$\mathbf{V!}$i	$\downarrow$	1
$\mathbf{V!}\pi$	$=$	$\rightarrow$	1
$=$	$\mathbf{N!}2$	$\rightarrow$	1
$\mathbf{N!}2$	$*$	$\uparrow$	1
$\mathbf{N!}2$	$\mathbf{M!}()2\text{x}1$	$\rightarrow$	1
$\mathbf{M!}()2\text{x}1$	$\mathbf{V!}$N	$\square$	1
$\mathbf{V!}$N	$\mathbf{V!}$i	$\multimap$	1

$$\pi_i = 2^{*}\binom{N}{i}$$



SLT Symbol Pair Generation

Window size = 1

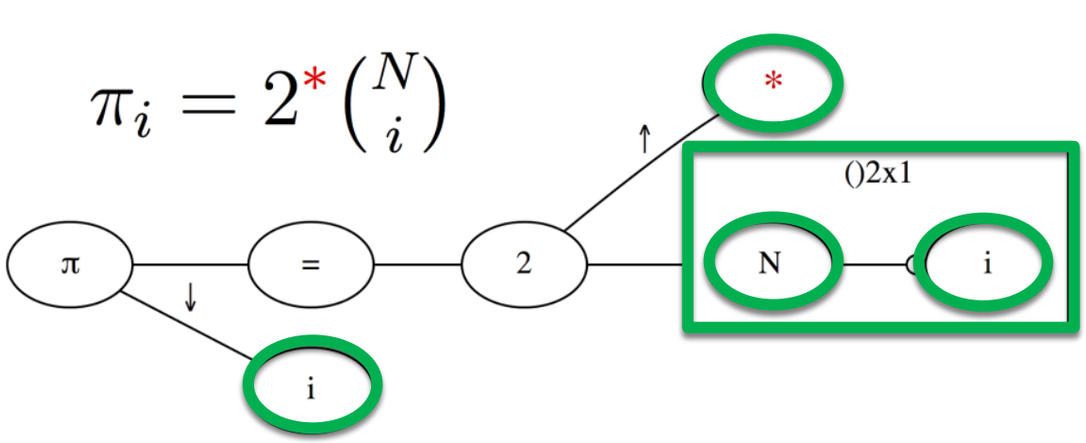


Sym-1	Sym-2	Path	Count
V!π	**V!**i	$\downarrow$	1
V!π	$=$	$\rightarrow$	1
$=$	**N!**2	$\rightarrow$	1
N!2	*	$\uparrow$	1
N!2	**M!**()2x1	$\rightarrow$	1
M!()2x1	**V!**N	$\square$	1
V!N	**V!**i	$\multimap$	1
V!N	!0	$\rightarrow$	1
V!i	!0	$\rightarrow$	2
M!()2x1	!0	$\rightarrow$	1
*	!0	$\rightarrow$	1

End-Of-Line = True

SLT Symbol Pair Generation

Window size = 1



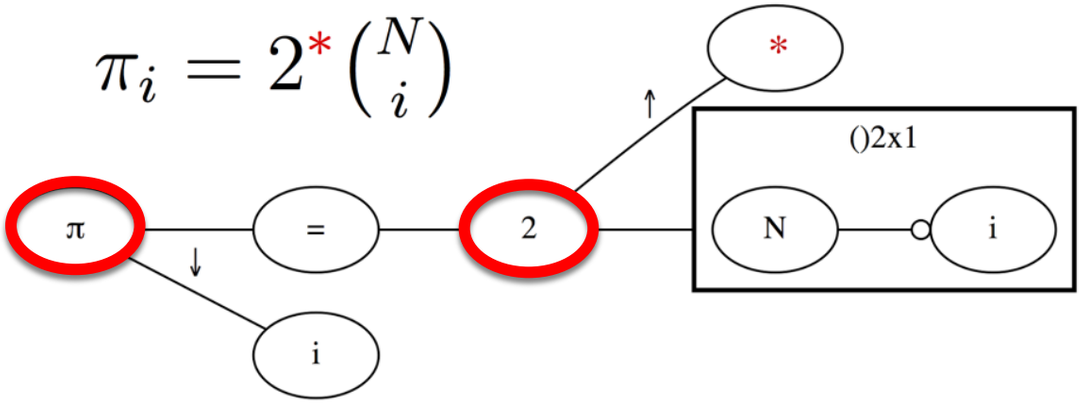
Sym-1	Sym-2	Path	Count
V!π	**V!**i	↓	1
V!π	=	→	1
=	**N!**2	→	1
N!2	*	↑	1
N!2	**M!**()2x1	→	1
M!()2x1	**V!**N	□	1
V!N	**V!**i	⊸	1
V!N	!0	→	1
V!i	!0	→	2
M!()2x1	!0	→	1
*	!0	→	1

End-Of-Line = True

SLT Symbol Pair Generation

Window size = 2

$$\pi_i = 2^{*}\binom{N}{i}$$

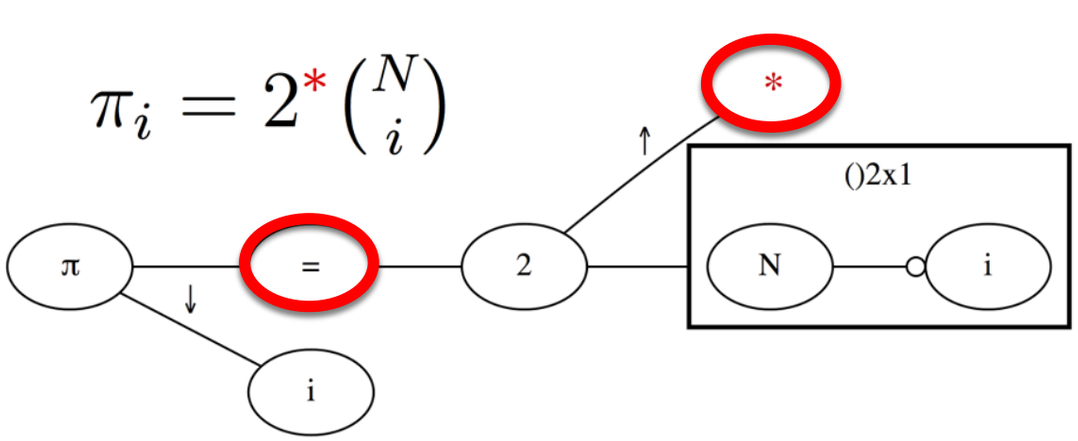


Sym-1	Sym-2	Path	Count
V!π	**V!**i	$\downarrow$	1
V!π	$=$	$\rightarrow$	1
$=$	**N!**2	$\rightarrow$	1
N!2	*	$\uparrow$	1
N!2	**M!**()2x1	$\rightarrow$	1
M!()2x1	**V!**N	$\boxed{\cdot}$	1
V!N	**V!**i	$\multimap$	1
V!N	!0	$\rightarrow$	1
V!i	!0	$\rightarrow$	2
M!()2x1	!0	$\rightarrow$	1
*	!0	$\rightarrow$	1
V!π	**N!**2	$\rightarrow\rightarrow$	1

SLT Symbol Pair Generation

Window size = 2

$$\pi_i = 2^{*}\binom{N}{i}$$

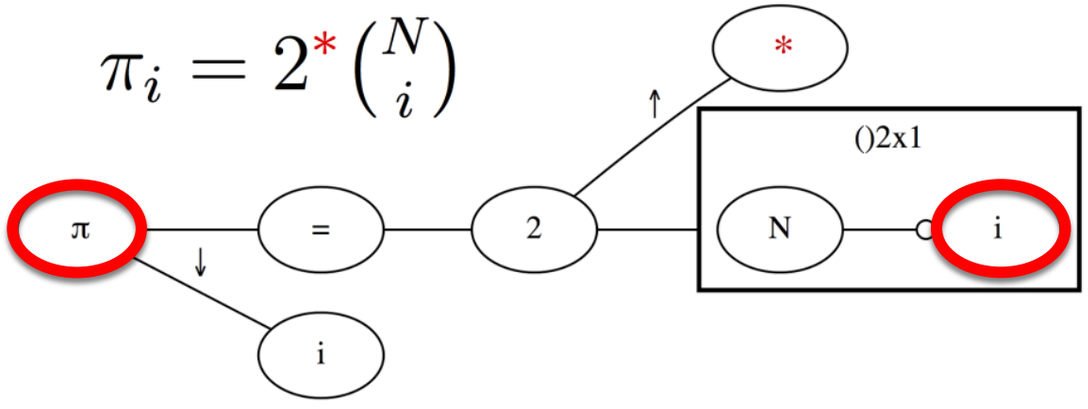


Sym-1	Sym-2	Path	Count
V!π	**V!**i	$\downarrow$	1
V!π	$=$	$\rightarrow$	1
$=$	**N!**2	$\rightarrow$	1
N!2	*	$\uparrow$	1
N!2	**M!**()2x1	$\rightarrow$	1
M!()2x1	**V!**N	$\square$	1
V!N	**V!**i	$\multimap$	1
V!N	!0	$\rightarrow$	1
V!i	!0	$\rightarrow$	2
M!()2x1	!0	$\rightarrow$	1
*	!0	$\rightarrow$	1
V!π	**N!**2	$\rightarrow\rightarrow$	1
$=$	*	$\rightarrow\uparrow$	1

SLT Symbol Pair Generation

Window size = All

$$\pi_i = 2^{*}\binom{N}{i}$$

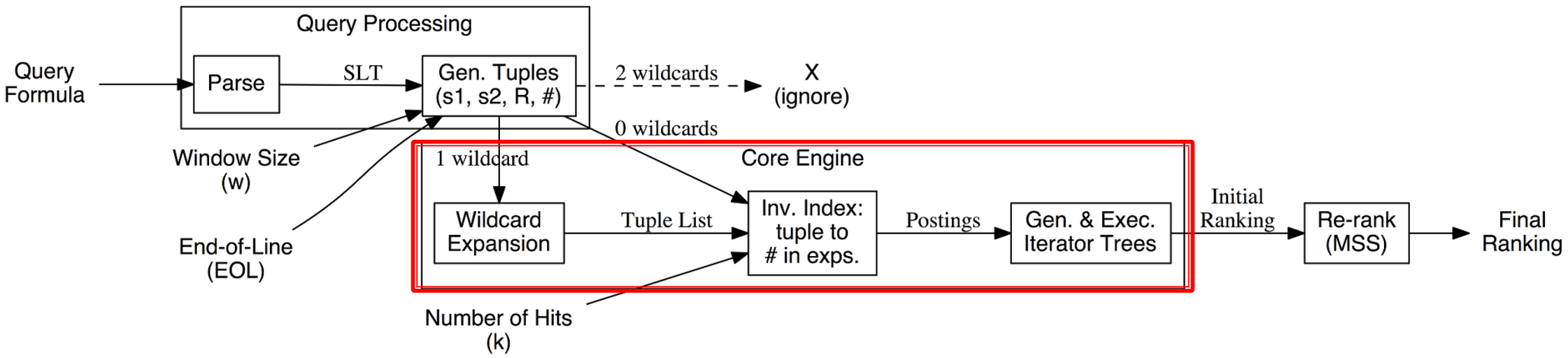


Sym-1	Sym-2	Path	Count
V!π	**V!**i	$\downarrow$	1
V!π	$=$	$\rightarrow$	1
$=$	**N!**2	$\rightarrow$	1
N!2	*	$\uparrow$	1
N!2	**M!**()2x1	$\rightarrow$	1
M!()2x1	**V!**N	$\boxed{\cdot}$	1
V!N	**V!**i	$\multimap$	1
V!N	!0	$\rightarrow$	1
V!i	!0	$\rightarrow$	2
M!()2x1	!0	$\rightarrow$	1
*	!0	$\rightarrow$	1
V!π	**N!**2	$\rightarrow\rightarrow$	1
$=$	*	$\rightarrow\uparrow$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
V!π	**V!**i	$\rightarrow\rightarrow\rightarrow\boxed{\cdot}\multimap$	1

Tangent-3 Formula Retrieval Model

Steps

1. **Query Processing:** symbol pair generation
2. **Core Engine:** symbol pair retrieval
3. **Re-rank:** Max. Subtree Similarity (MSS)



Core Engine

Inverted Index for Symbol Pairs

Entries: Pairs of symbols with relationships (s_1, s_2, R)

Posting lists: Formulas containing pairs with counts (#)

Document Index: Maps formulas to documents containing them

Fast query evaluation using Iterator trees

Wildcards iterate over multiple posting lists

Formulas scored using Dice coefficient of pairs

Optimizations for faster retrieval

$$S = \frac{2|Q \cap C|}{|Q| + |C|}$$

Q = Query Pairs C = Candidate Pairs

Core Engine – Query Evaluation

Query

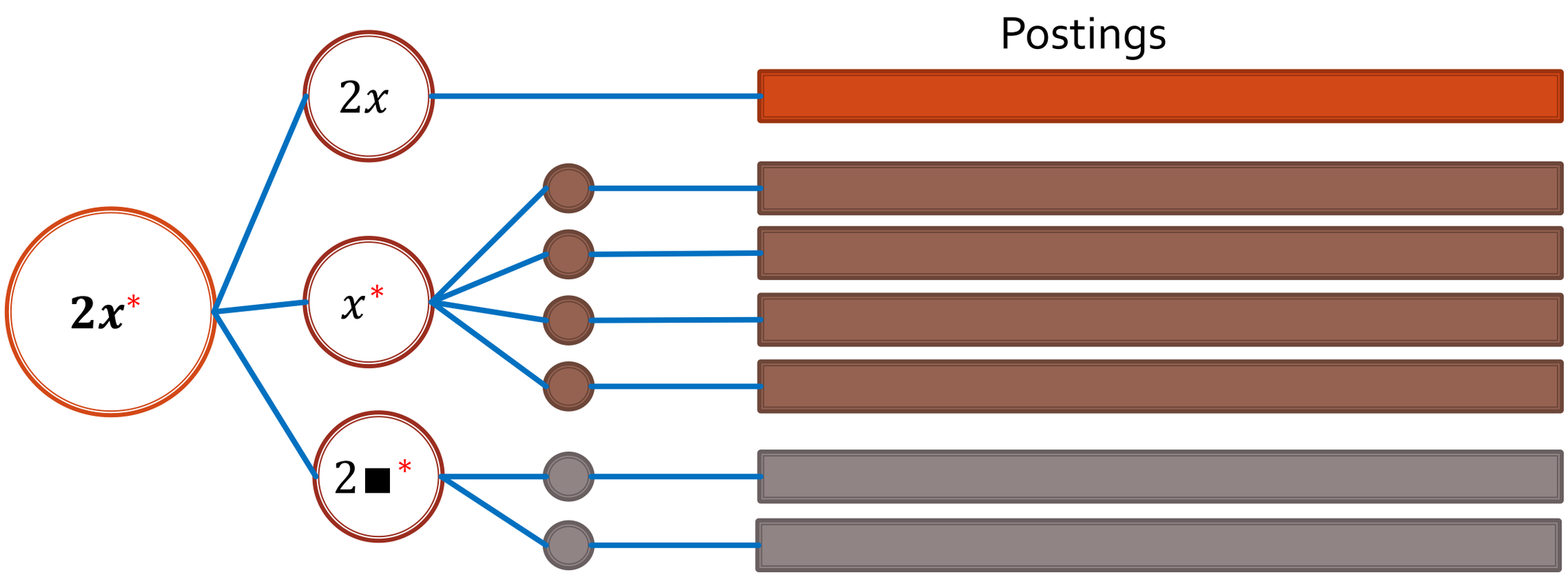
$$2x^*$$

Query pairs

Sym-1	Sym-2	Path	Count
2	X	→	1
X	*	↑	1
2	*	→ ↑	1

Query Evaluation – Iterator Tree

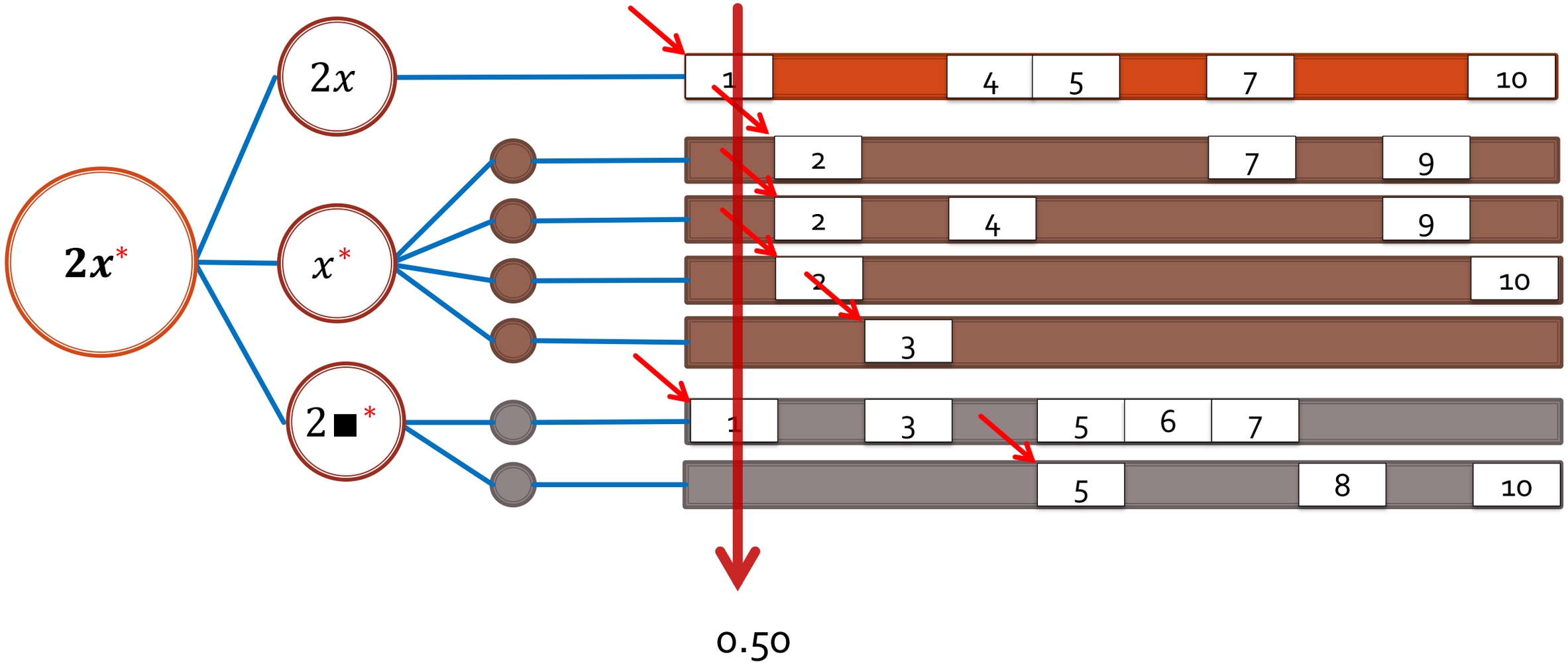
Sym-1	Sym-2	Path	Count
2	X	→	1
X	*	↑	1
2	*	→ ↑	1



Iterator Tree Evaluation

Priority Queue (Form. IDs + Score)

Iterator Tree Evaluation

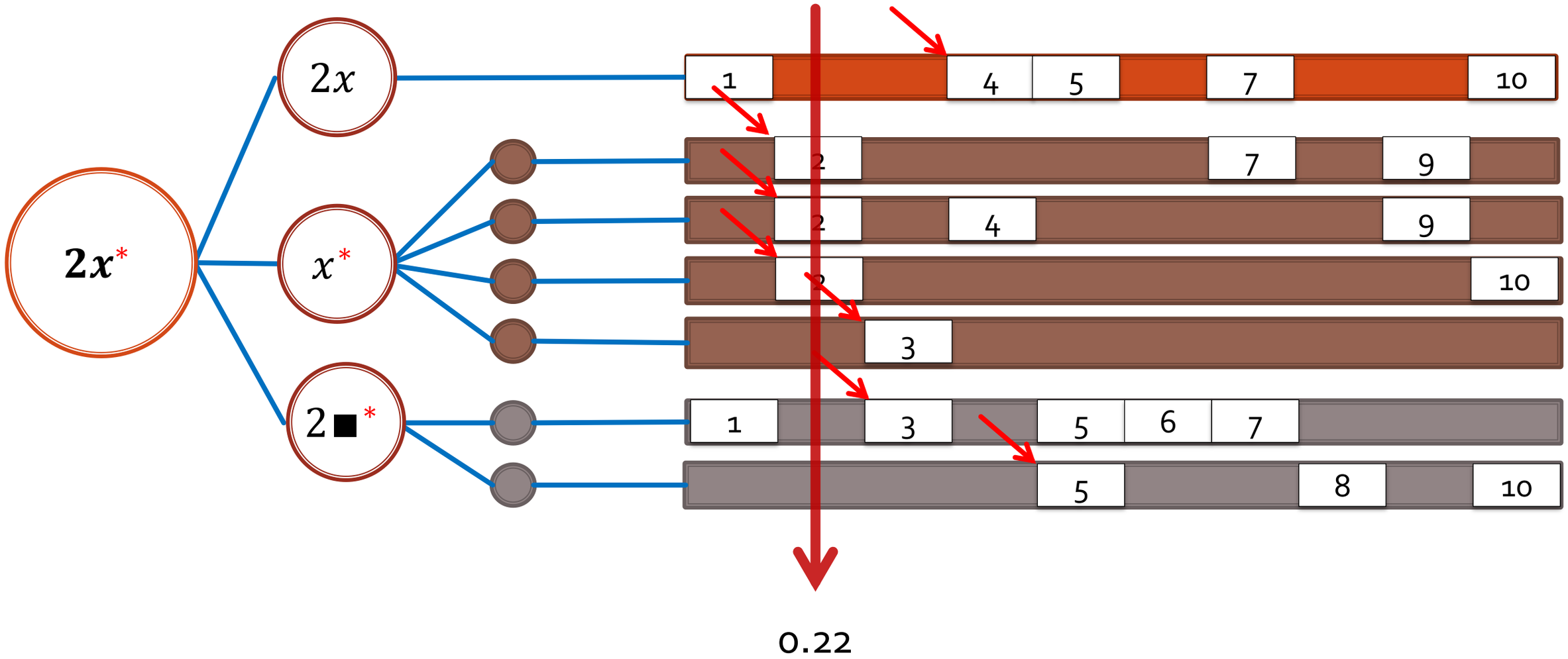


Priority Queue (Form. IDs + Score)

1

0.50

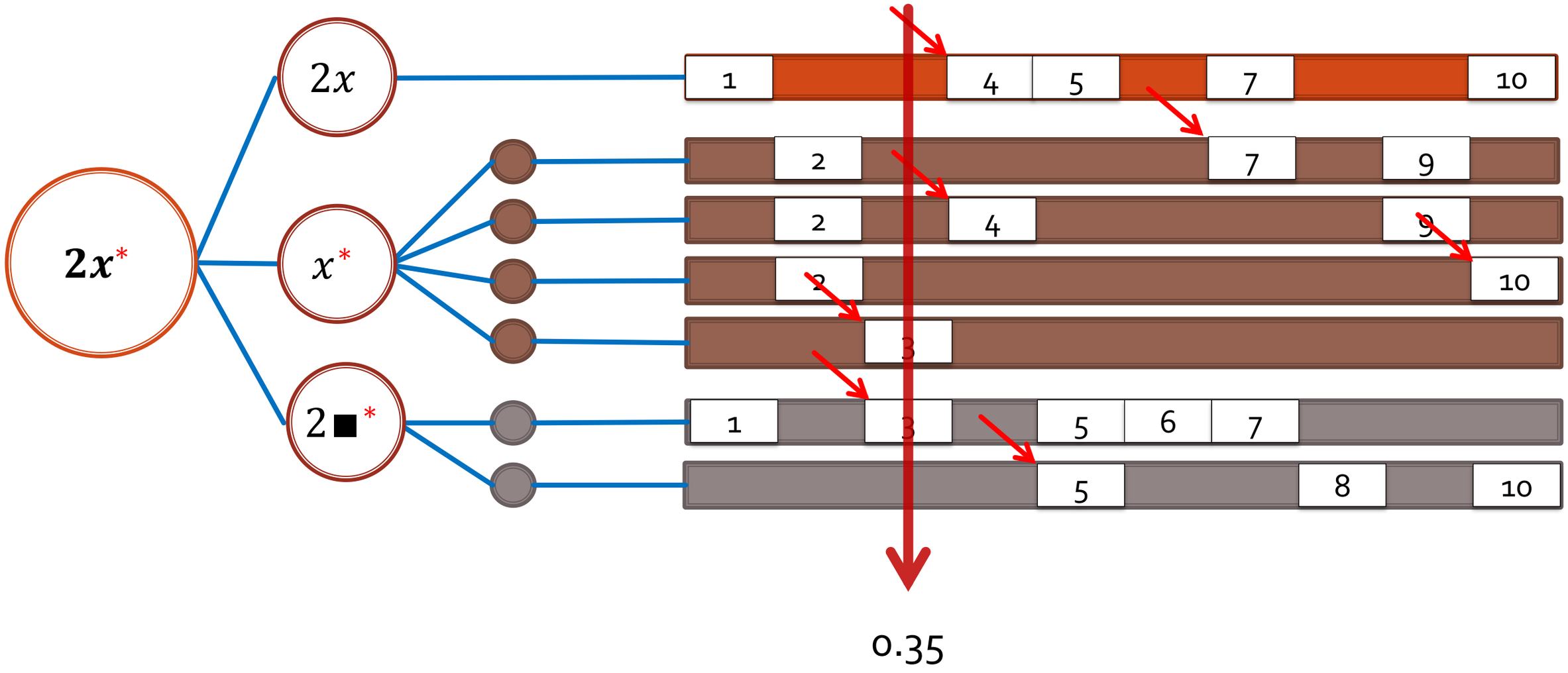
Iterator Tree Evaluation



Priority Queue (Form. IDs + Score)

1	2
0.50	0.22

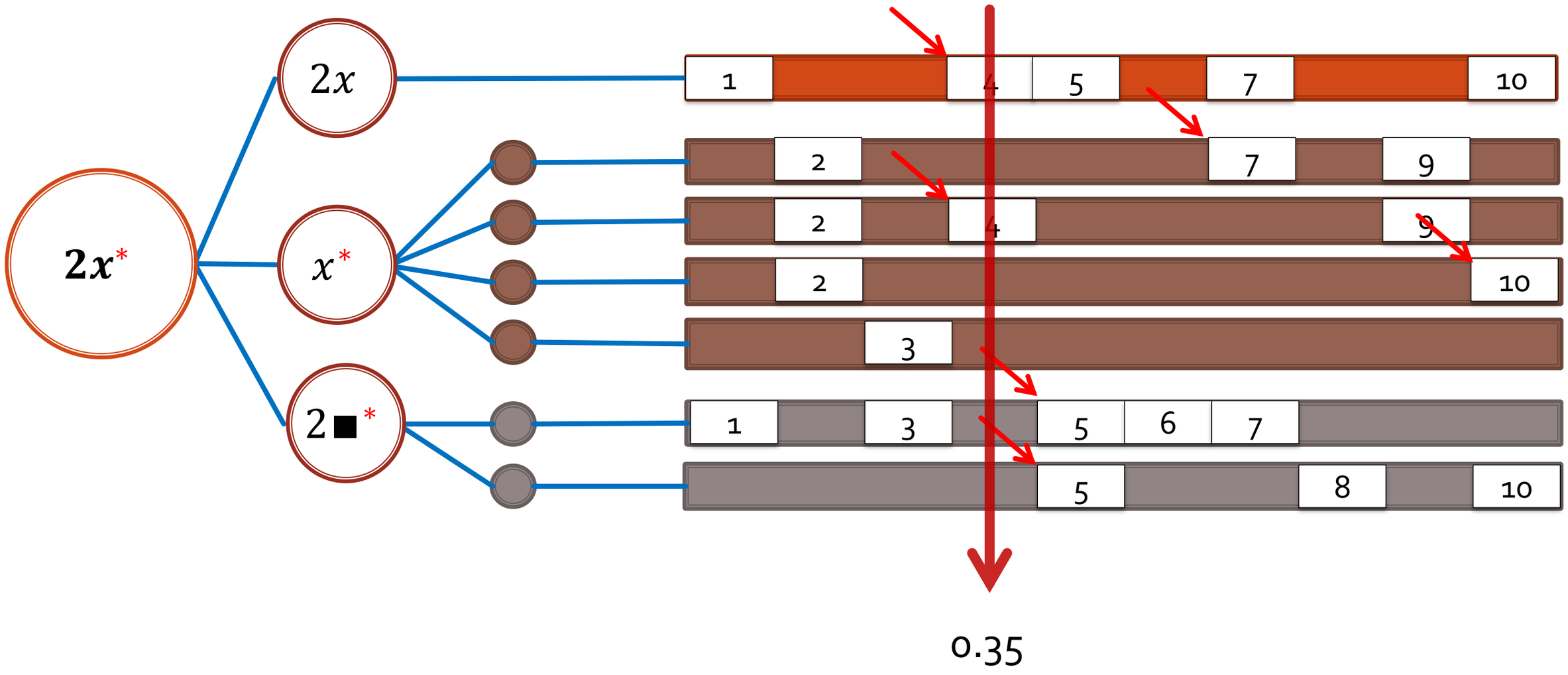
Iterator Tree Evaluation



Priority Queue (Form. IDs + Score)

1	3	2
0.50	0.35	0.22

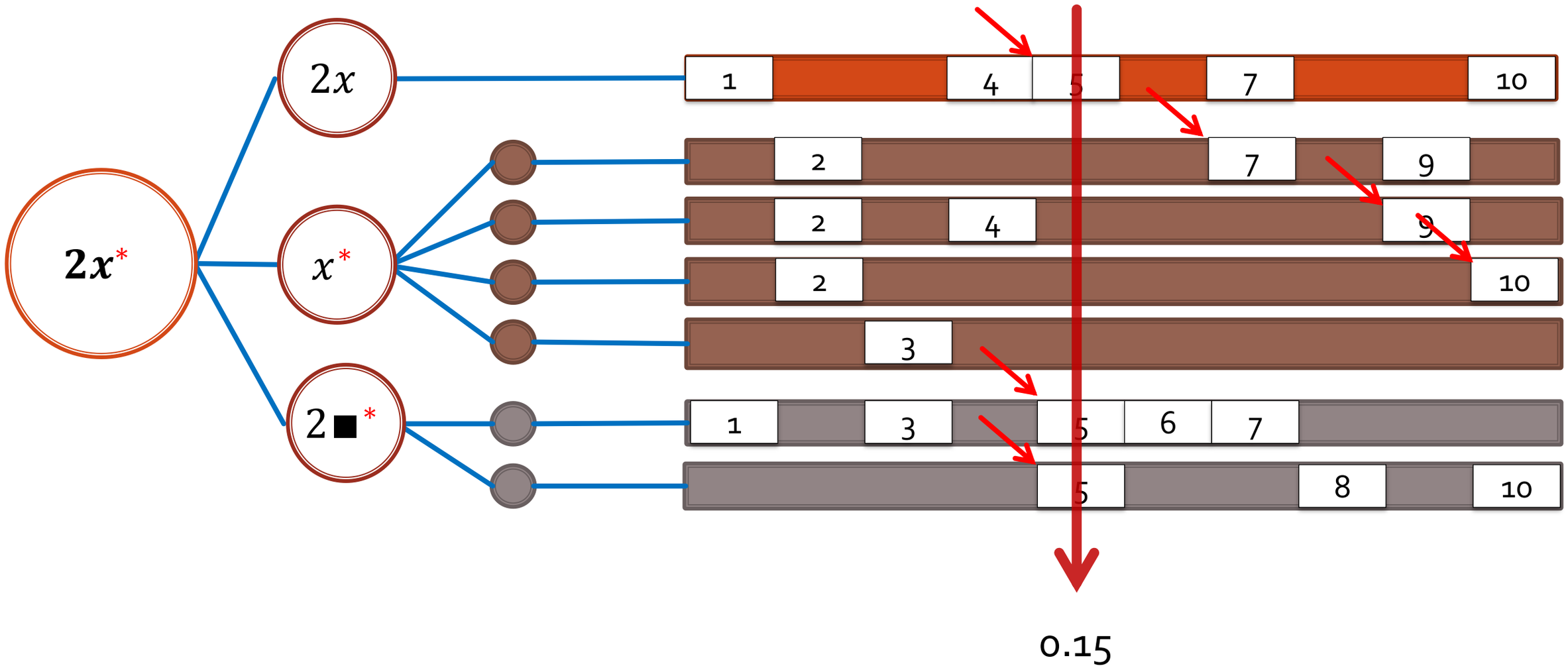
Iterator Tree Evaluation



Priority Queue (Form. IDs + Score)

1	3	4	2
0.50	0.35	0.35	0.22

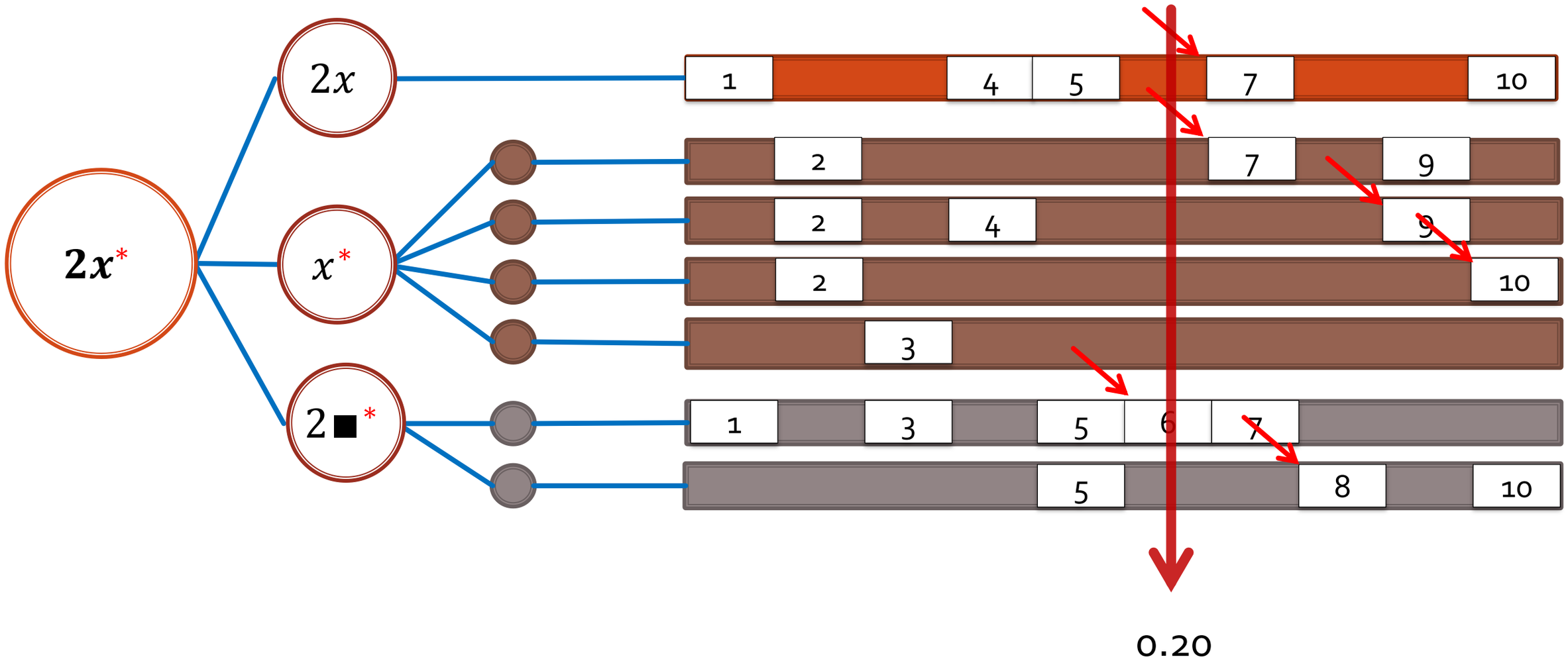
Iterator Tree Evaluation



Priority Queue (Form. IDs + Score)

1	3	4	2	5
0.50	0.35	0.35	0.22	0.15

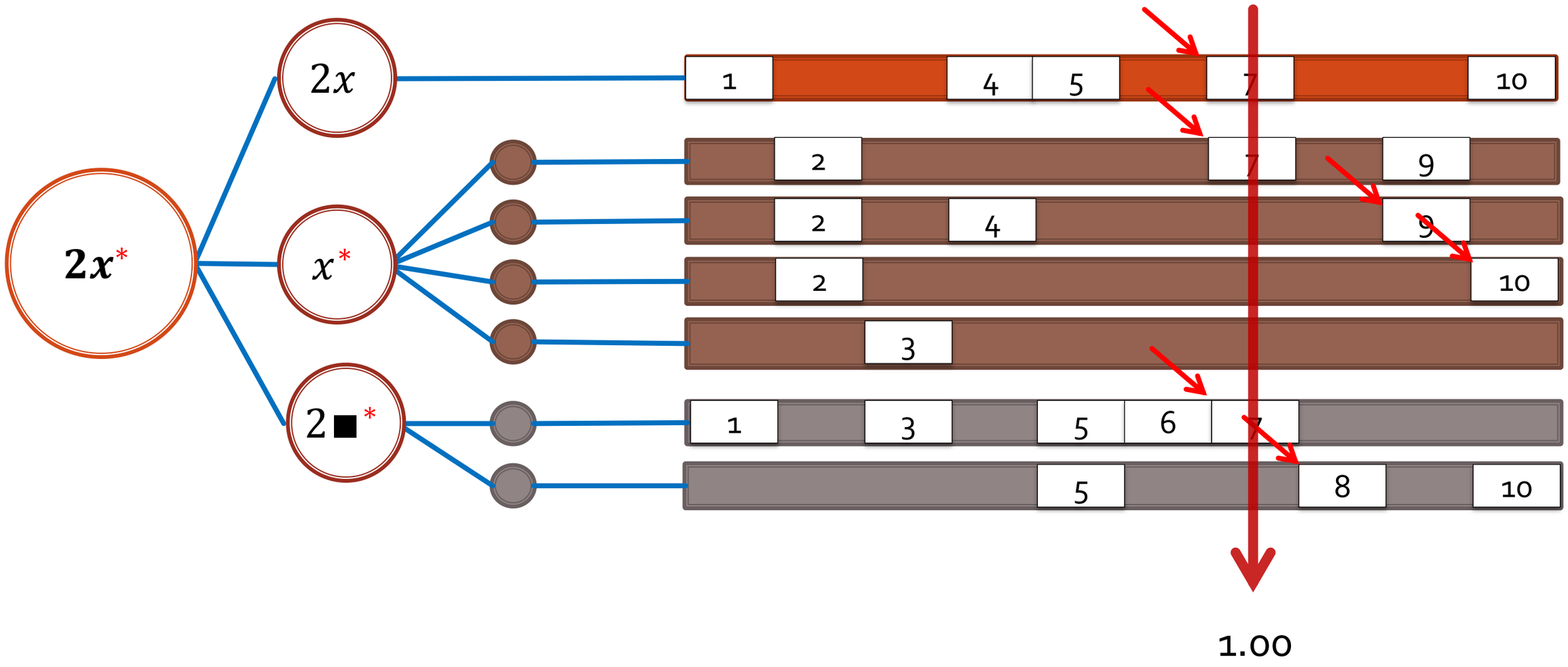
Iterator Tree Evaluation



Priority Queue (Form. IDs + Score)

1	3	4	2	6	5
0.50	0.35	0.35	0.22	0.20	0.15

Iterator Tree Evaluation



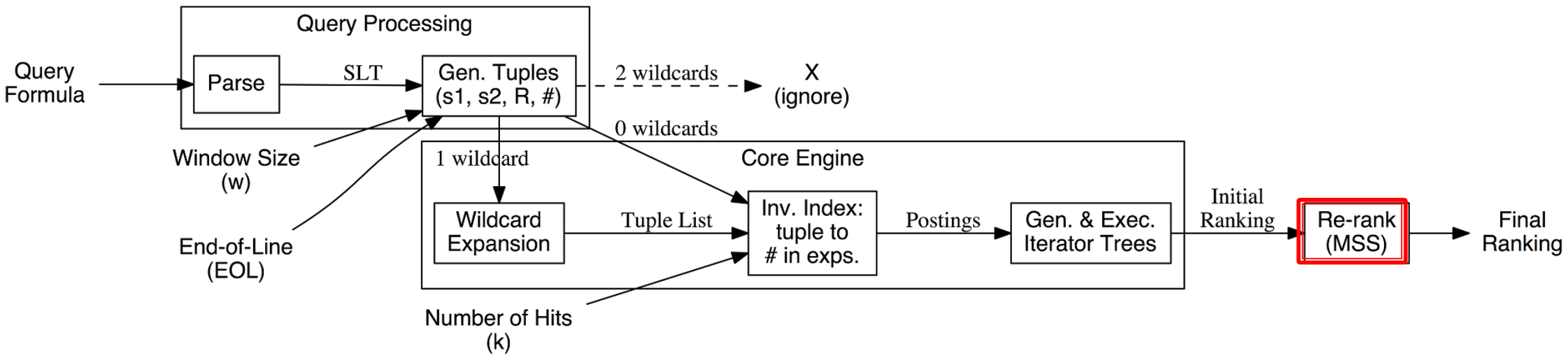
Priority Queue (Form. IDs + Score)

7	1	3	4	2	6	5
1.00	0.50	0.35	0.35	0.22	0.20	0.15

Tangent-3 Formula Retrieval Model

Steps

1. **Query Processing:** symbol pair generation
2. **Core Engine:** symbol pair retrieval
3. **Re-rank:** Max. Subtree Similarity (MSS)



Re-ranking Example 1

QUERY 1: $f_*(z) = z^2 + c$

	INITIAL RANKING	RE-RANKED (MSS)
1.	$f_c(z) = z^2 + c$	$f_c(z) = z^2 + c$
2.	$f_c(z) = z^2 + c.$	$\mathbf{P}_c(z) = z^2 + c$
3.	$f(z) = z^2 + c$	$f_c(\mathbf{x}) = \mathbf{x}^2 + c$
4.	$f_0(z) = z^2$	$f_c(z) = z^2 + c.$
5.	$f_c(z) = z * z + c$	$f(z) = z^2 + c$

***Add unification of symbols for top-k results;**
Core *partially* unifies wildcards only

Re-ranking Example 2

QUERY 2: $\sum_{*2*}^{*1*} * = \sum_{*2*}^{*1*} *$

	INITIAL RANKING	RE-RANKED (MSS)
1.	$E = \sum_i^N E_i$	$\sum_{i=1}^d a_i = \sum_{i=1}^d b_i$
2.	$G_{net} = \sum_i \sum_{i=1}^N$	$\sum_{i=1}^N d_i = \sum_{i=1}^N \lambda_i .$
3.	$\sum_i^{N_1} p_i = \sum_j^{N_2} p_j$	$\sum_{n=0}^{\infty} a_{\sigma(n)} = \sum_{n=0}^{\infty} a_n .$
4.	$\sum_{i=1}^n x_i k_i = \sum_{i=1}^n x_i$	$\sum_i^{N_1} p_i = \sum_j^{N_2} p_j$
5.	$= \sum_{k=1}^n a_k$	$\sum_{n=0}^{\infty} a_n = \sum_{n \in N} a_n .$

***Use SLT structure directly, and support wildcard constraints;**
Core performs spectral pair matching, no wildcard constraints (for speed)

Re-ranking

Re-scores top-k candidates using a triple

Steps:

1. Locate and score Query matches on Candidate
2. Select best match

Detailed SLT matching (slower than Core – $O(n^3 \log n)$)

- Add unification of symbols for top-k results
- Use SLT structure directly
- Support wildcard constraints

Unification of Symbols

Unifiable Types: Identifier, Number, Matrix/Group, Wildcard

Query: $y = \frac{1}{2}\left(9.8\,{}^{m}/_{s^2}\right)t^2$

$y = \frac{1}{2}\left(9.807\,{}^{m}/_{s^2}\right)t^2$

Candidate 1

$h = \frac{1}{2}\left(32.174\,{}^{ft}/_{s^2}\right)t^2$

Candidate 2

Maximum Subtree Similarity (MSS)

Based on

Dice coefficient of query symbol (R_s) and relationship (R_p) recall

Tie breakers

Extra symbols (-)

Exactly matched symbols

$$R_s = \frac{|E \cup U \cup W_q|}{|T_q|}$$

$$R_p = \frac{Edges(E \cup U \cup W_q)}{Edges(T_q)}$$

$$MSS = \frac{2R_pR_s}{R_p + R_s}$$

$$Score = (MSS, -U, E)$$

$E =$ Exact Matches

$U =$ Unified Matches

$W_q =$ Wildcards Matched

$T_q =$ Query SLT

MSS ordering example

Query: $S(k)$

$$S(k) \;>\; F(k) \;>\; F(k) + S(k) \;>\; (k) \;>\; S^{(k)}$$

$(1, 0, 3)$ $(1, 0, 2)$ $(1, -4, 3)$ $(0.6, 0, 2)$ $(0.6, -1, 2)$

Scoring Triples:

1st element: MSS Query structure match score

2nd element: Unmatched symbols in Query (negative)

3rd element: Identical symbols in Query and candidate

Sort triples in descending lexicographic order.

Results

Experiments: Effectiveness & Efficiency

System: Ubuntu Linux 14.04, 24 x 2.93 GHz Intel proc., 96 GB RAM

Queries: 100 NTCIR-11 Wikipedia formula queries

All results are for k=100 formulae returned from the Core

Corpora

1. **NTCIR-11 Wikipedia** (2.5 GB) 30,000 articles
   - 387,947 LaTeX expressions
2. **NTCIR-11 arXiv** (174 GB) 8,301,578 excerpts (~paragraphs)
   - 60 million formulae (incl. isolated symbols)

NTCIR-11 Wikipedia Results

Task: Retrieve formula at specific location in one document
Results: Top 10,000 results (**Tangent-3:** expand k=100 results to docs.)
Const (65): no wildcards **Var (35):** wildcards

		Recall@k					
		Documents			Formulae		
System		Total	Const	Var	Total	Const	Var
TUW Vienna †		**97**	*100	91	**93**	98	83
NII Japan †		**97**	98	*94	**94**	97	89
Tangent-2 ∘		**88**	91	83	**78**	78	77
Tangent-3 Core ∘							
w=1	No-EOL	**95**	97	91	**95**	97	91
	Sm-EOL	**97**	*100	91	**97**	*100	91
	EOL	***98**	*100	*94	***98**	*100	*94
w=All	No-EOL	**95**	97	91	**95**	97	91
	Sm-EOL	**97**	*100	91	**97**	*100	91
	EOL	**97**	*100	91	**96**	*100	89

†: uses Operator Tree (OT) formula representation (Content MathML)
∘: uses Symbol Layout Tree (SLT) formula representation

Tangent-3

Highest formula recall@k
all Const. for (Sm-EOL, EOL)

Highest doc. recall@k
match (Sm-EOL) or exceed (EOL +1 Var.)

Larger w: no benefit

Re-ranking: no effect

NTCIR-11 Wikipedia Results

Task: Retrieve formula at specific location in one document
Results: Top 10,000 results (**Tangent-3:** expand k=100 results to docs.)
Const (65): no wildcards **Var (35):** wildcards

		MRR					
		Documents			Formulae		
System		Total	Const	Var	Total	Const	Var
TUW Vienna †		**80**	80	79	***82**	86	75
NII Japan †		**74**	79	74	**72**	*87	63
Tangent-2 ∘		**70**	68	75	**67**	65	72
Tangent-3 Core ∘							
w=1	No-EOL	**79**	80	77	**76**	76	76
	Sm-EOL	**80**	*82	77	**77**	78	76
	EOL	**81**	*82	79	**77**	78	76
w=All	No-EOL	**79**	80	78	**76**	76	76
	Sm-EOL	**80**	*82	78	**78**	78	76
	EOL	**81**	*82	78	**77**	78	75
Tangent-3 Re-rank ∘							
w=1	No-EOL	**80**	80	*82	**77**	76	*80
	Sm-EOL	***82**	*82	*82	**79**	78	*80
	EOL	***82**	*82	*82	**79**	78	*80
w=All	No-EOL	**80**	80	80	**77**	76	77
	Sm-EOL	**81**	*82	80	**78**	78	77
	EOL	***82**	*82	*82	**78**	78	78

MRR
Mean Reciprocal Rank (1/r)
Avg. at Rank 2 = 50%

Tangent-3

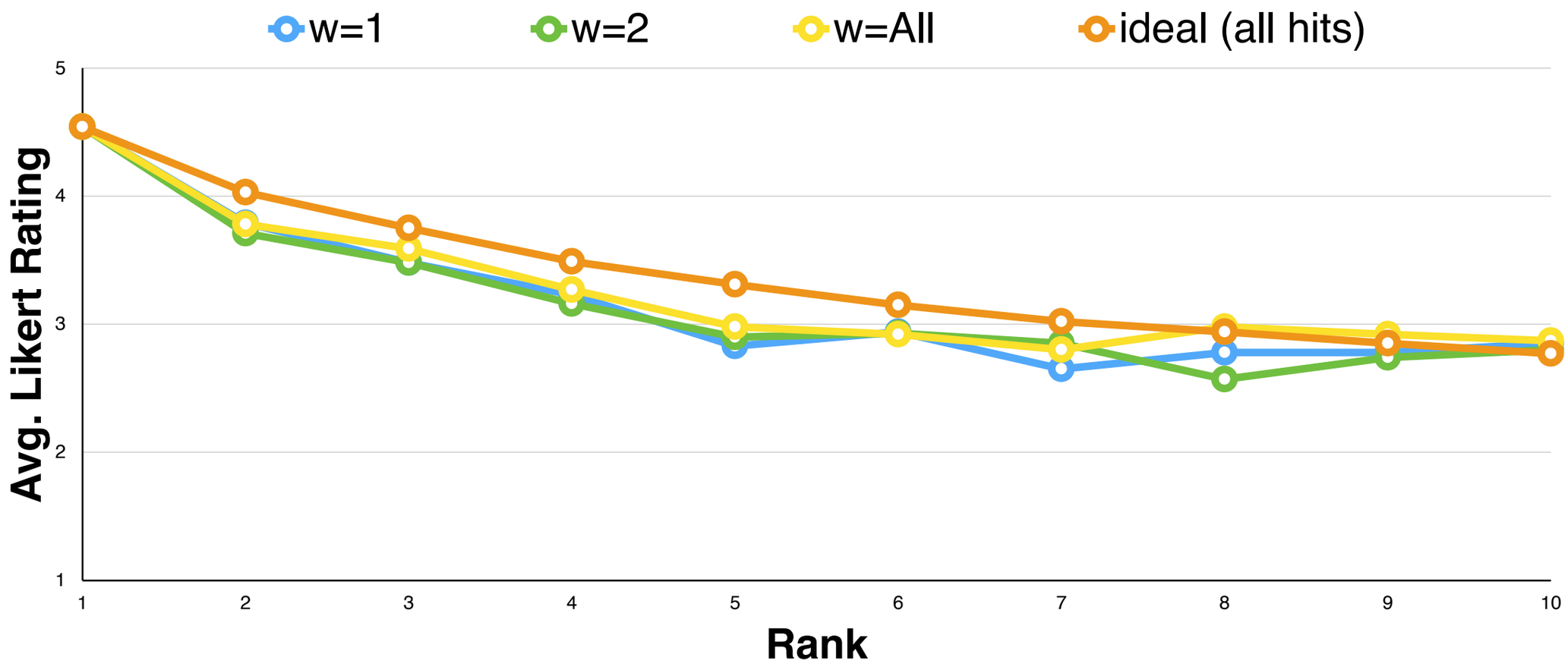
Highest doc. MRR *w. Rerank*

Highest Var MRR *w. Rerank*

Large w: little/no benefit

Lower Const. MRR
multiple identical matches

MSS Effect on Top-10 Similarity Ratings: Human Ratings for 10 Wiki Queries (*No-EOL*)



Ratings: 1: Very Dissimilar, **2:** Dissimilar, **3:** Neutral, **4:** Similar, **5:** Very Similar

21 participants from Computing College at RIT; evaluated pooled **Top-10 hits**

Randomized presentation, one hit-at-a-time (randomize queries, then hits for queries)

Similar results across window sizes

4 of the Top-5 hits *identical* for all but one query

Wikipedia Retrieval Times
(100 Queries, *k=100*)

Re-rank times consistent across window sizes, EOL entries

μ = 0.78 secs. σ = 3.56 secs. **median** = 0.07 secs.

Re-ranker in Python; C/C++ would accelerate substantially

Tangent-2 (9 parallel indices w. Amazon web services)

> 8 minutes (480 secs.)

Tangent-3 (Single process)

(w=1, No-EOL) **Core:** 0.57 secs + **Rerank:** 78.03 secs

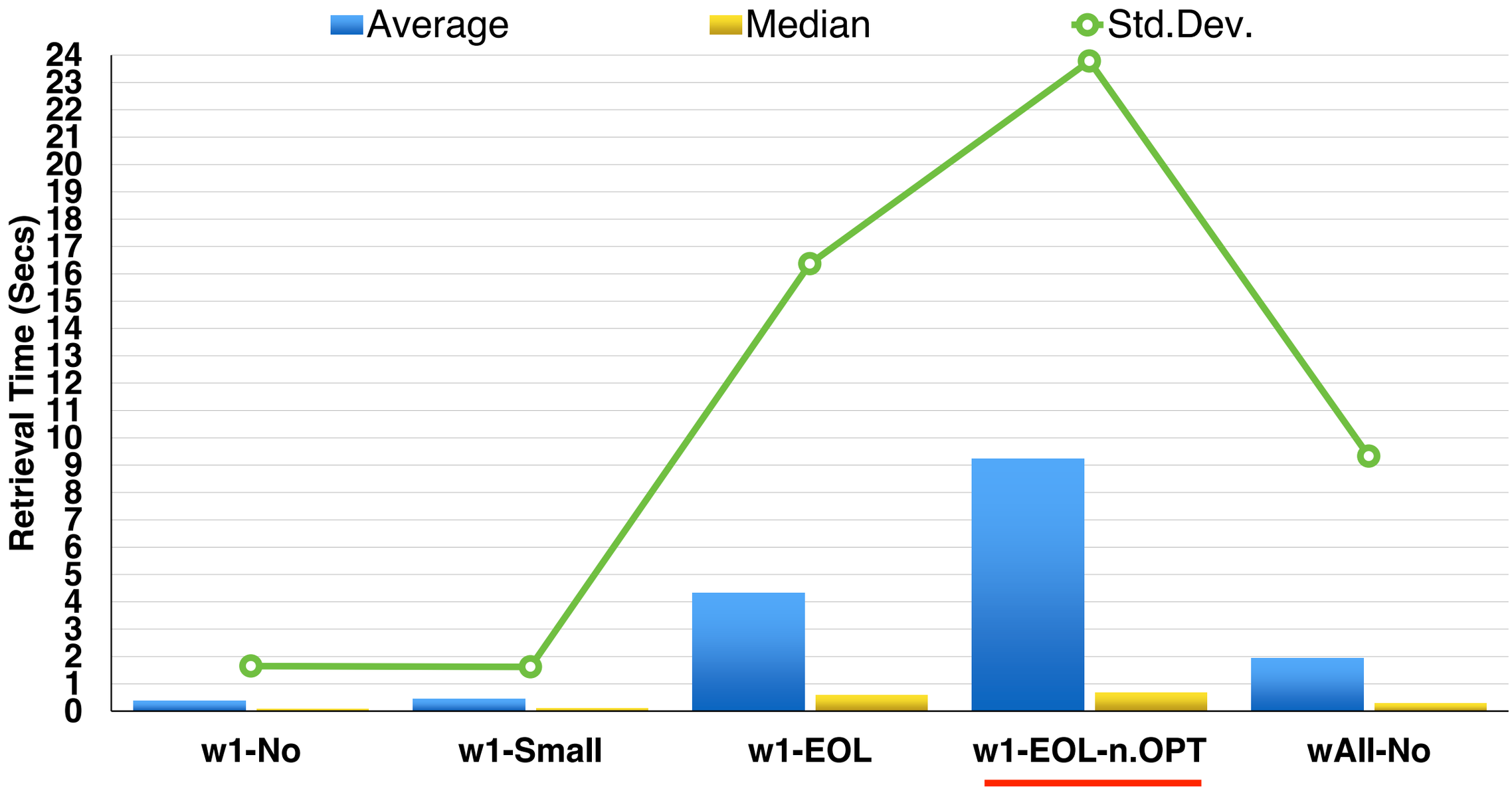
(w=All, EOL) **Core:** 8.48 secs + **Rerank:** 106.14 secs

Outlier: Rerank of 46 seconds (1.47 secs out of Core: 16 wildcards)

arXiv Core Retrieval Times
(100 Queries, *k=100*)

Window size (w) affects speed less than end-of-line entries (EOL)

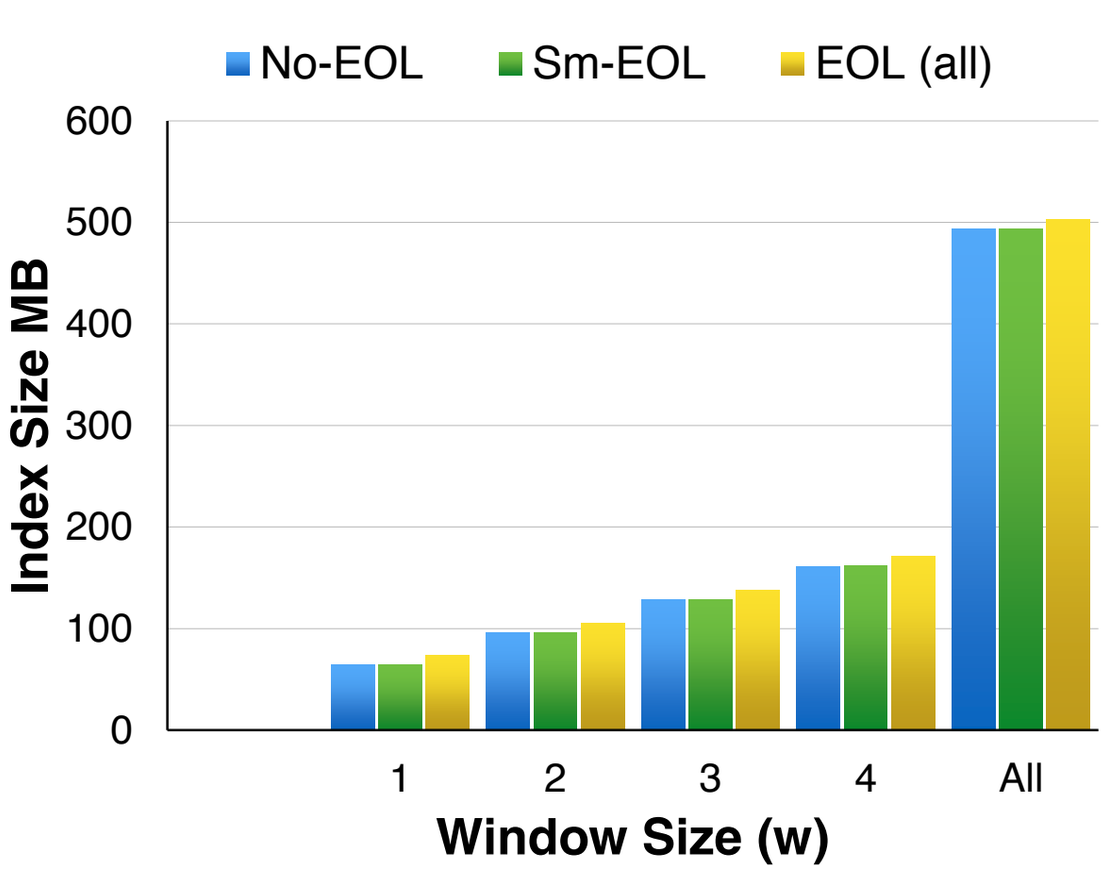
Core optimizations reduce variance

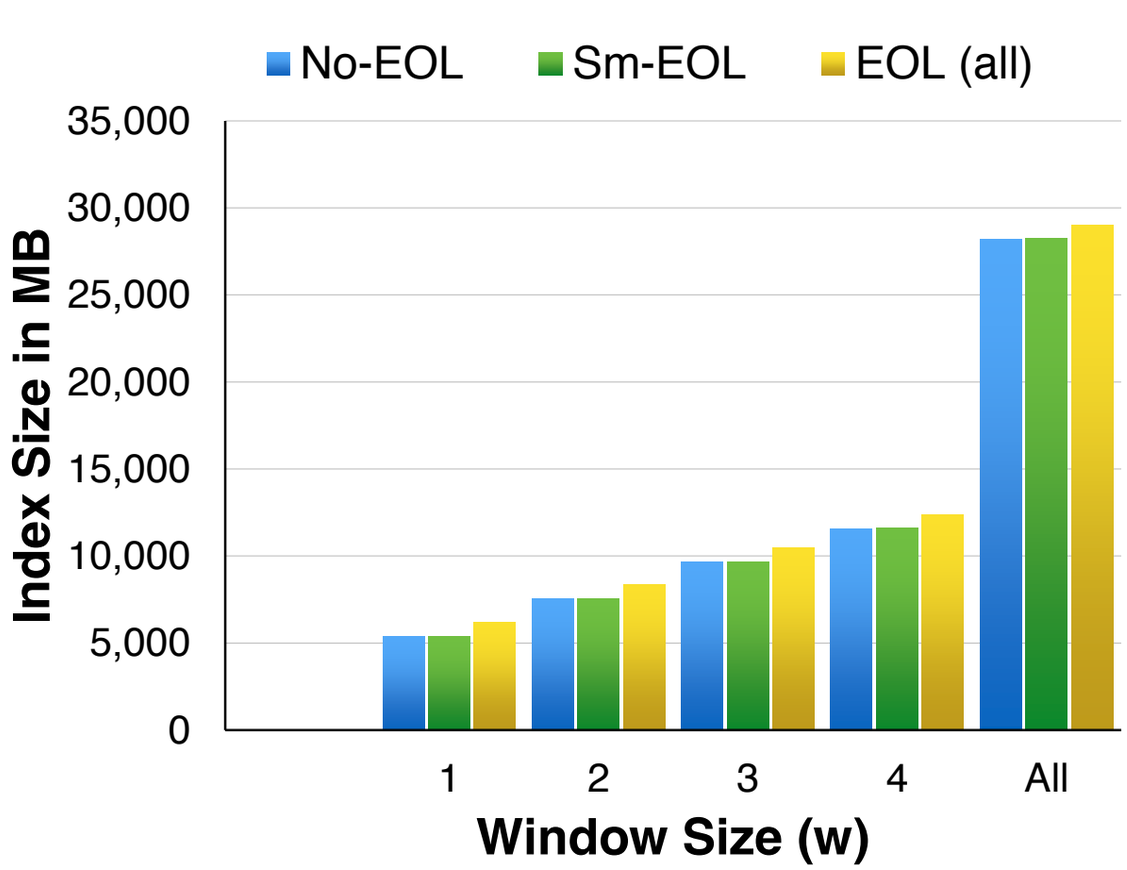


Index Sizes (on Disk)

Tangent-2: 1.3 GB (Wiki) and 36 GB (arXiv)

Tangent-3: *all* indices smaller than Tangent-2





**Indices 2-2.5 times larger when loaded into memory*

Conclusion

Tangent-3: two-stage appearance-based formula search engine that is **fast, effective and scalable**

Maximum Subtree Similarity (MSS): Dice coefficient for Query node + edge match

- Detailed structural matching, does not unfairly penalize exact query matches in large expressions
- Better supports symbol unification and wildcard constraints

Future Work

Adapt for **Operator Tree** representations - combine w. appearance?

Other notations (music, chemistry, etc.) / graphics

Additional Core accelerations

Improve MSS

- Speed; wildcard support; unification; match candidates

Thank you.

Source code: www.cs.rit.edu/~dprl/Software.html

We thank SIGIR and the ACM for providing a student travel award.

This material is based upon work supported by the **National Science Foundation** (USA) under Grants No. IIS-1016815 and HCC-1218801.

Financial support from the **Natural Sciences and Engineering Research Council of Canada** under Grant No. 9292/2010, **Mitacs**, and the **University of Waterloo** is gratefully acknowledged.