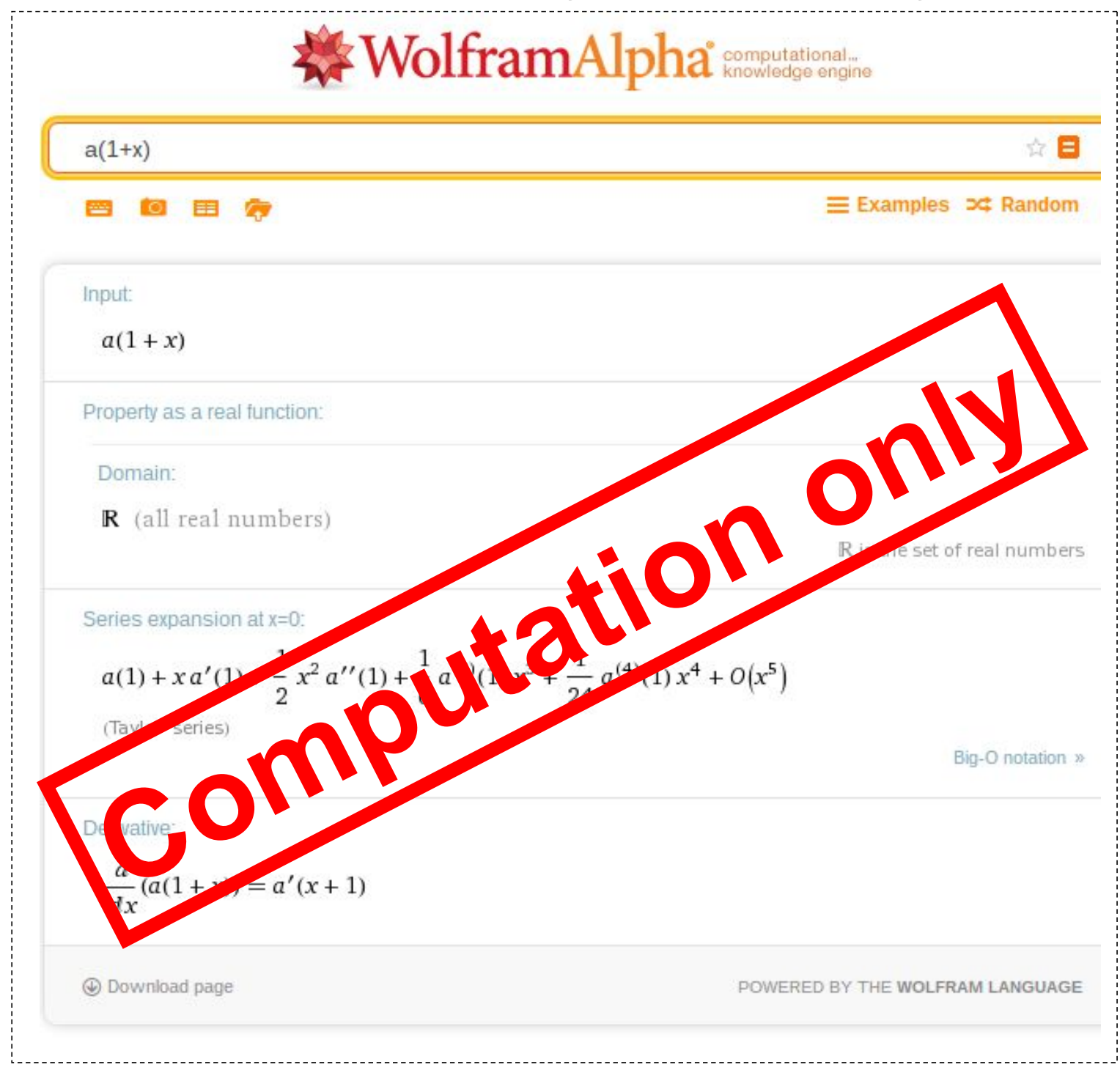
OPMES: A similarity search engine for mathematical content

Wei Zhong and Hui Fang
University of Delaware, USA

Motivation

Popular search engines are **unable** to search math expressions by similarity.



WolframAlpha computational knowledge engine

a(1+x)

Examples Random

Input:

$a(1+x)$

Property as a real function:

Domain:

ℝ (all real numbers)

ℝ is the set of real numbers

Series expansion at x=0:

$a(1)+x\,a'(1)+\frac{1}{2}x^2\,a''(1)+\frac{1}{6}a^{(3)}(1)\,x^3+\frac{1}{24}a^{(4)}(1)\,x^4+O(x^5)$

(Taylor series)

Big-O notation »

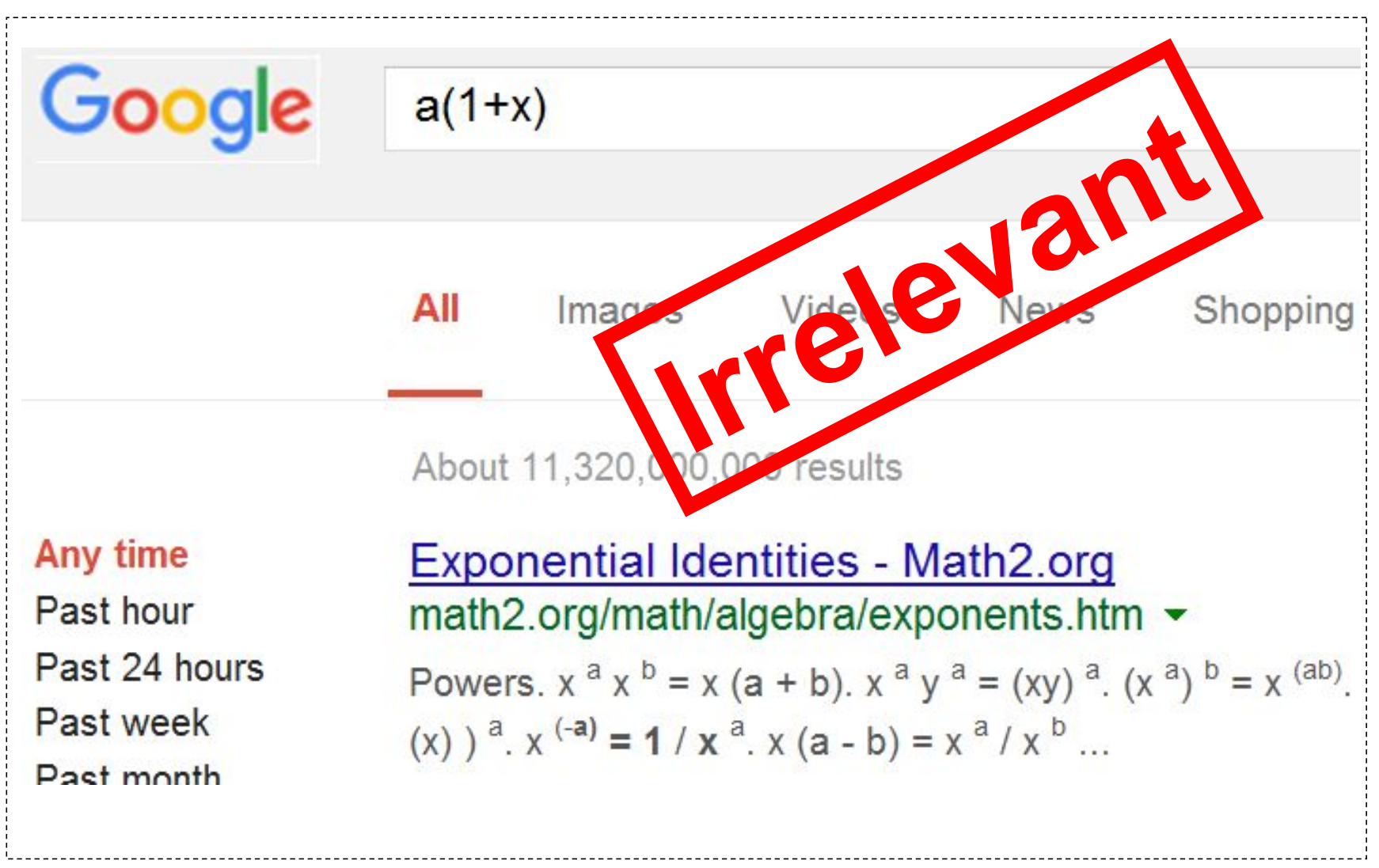
Derivative:

$\frac{d}{dx}(a(1+x))=a'(x+1)$

Download page

POWERED BY THE WOLFRAM LANGUAGE

Computation only



Google

a(1+x)

All Images Videos News Shopping

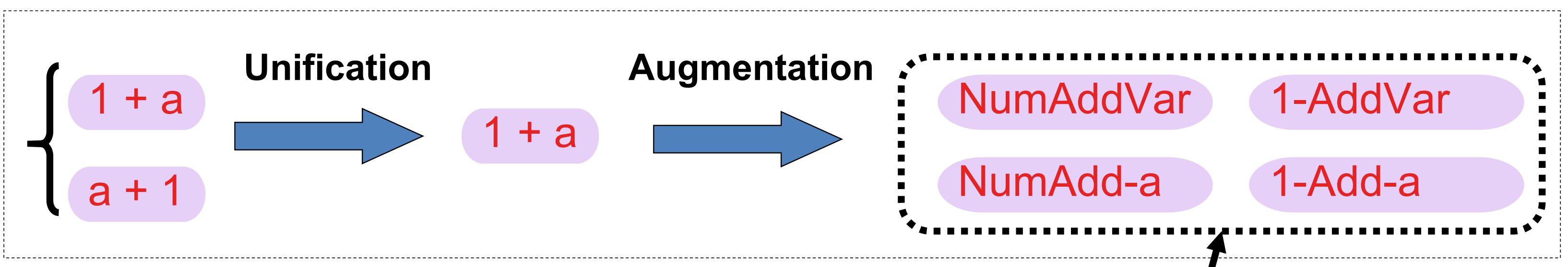
About 11,320,000,000 results

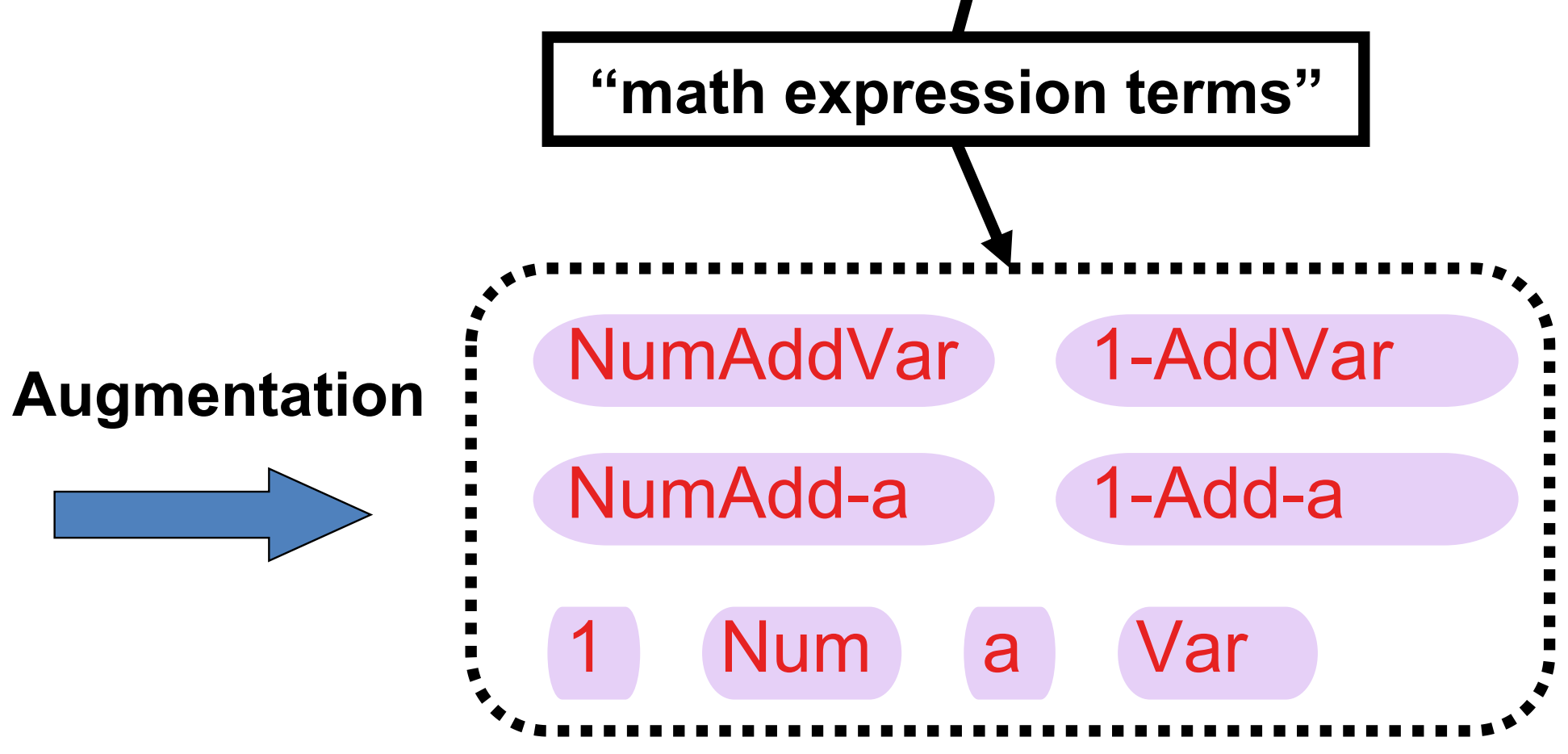
Any time
Past hour
Past 24 hours
Past week
Past month

Exponential Identities - Math2.org
math2.org/math/algebra/exponents.htm
Powers. x a x b = x (a + b). x a y a = (xy) a. (x a) b = x (ab). (x)) a. x (-a) = 1 / x a. x (a - b) = x a / x b ...

Irrelevant

More researchers are focusing on math similarity search. There are huge research potential to improve, especially in bringing novel ideas so that we avoid the fundamental drawbacks of existing text-only model/tools. e.g. text-based methods inevitably requires complicated unification process and large storage space as expressions are frequently augmented:





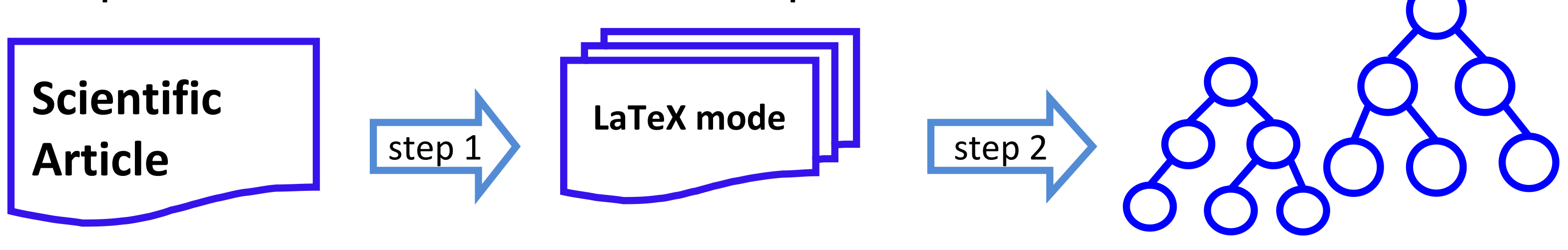
To further search math expressions by its subexpressions, we need even more space at the stage of augmentation.

Parsing LaTeX into Operator Tree

We use a tree-based method to remove unnecessary augmentation. The fundamental difference between tree-based approach and text-based math similarity search method is the former one generates an in-memory (intermediate) tree to extract structural information of math expressions.

We choose to convert math formula into operator tree and the way we are doing this is using a LALR parser :

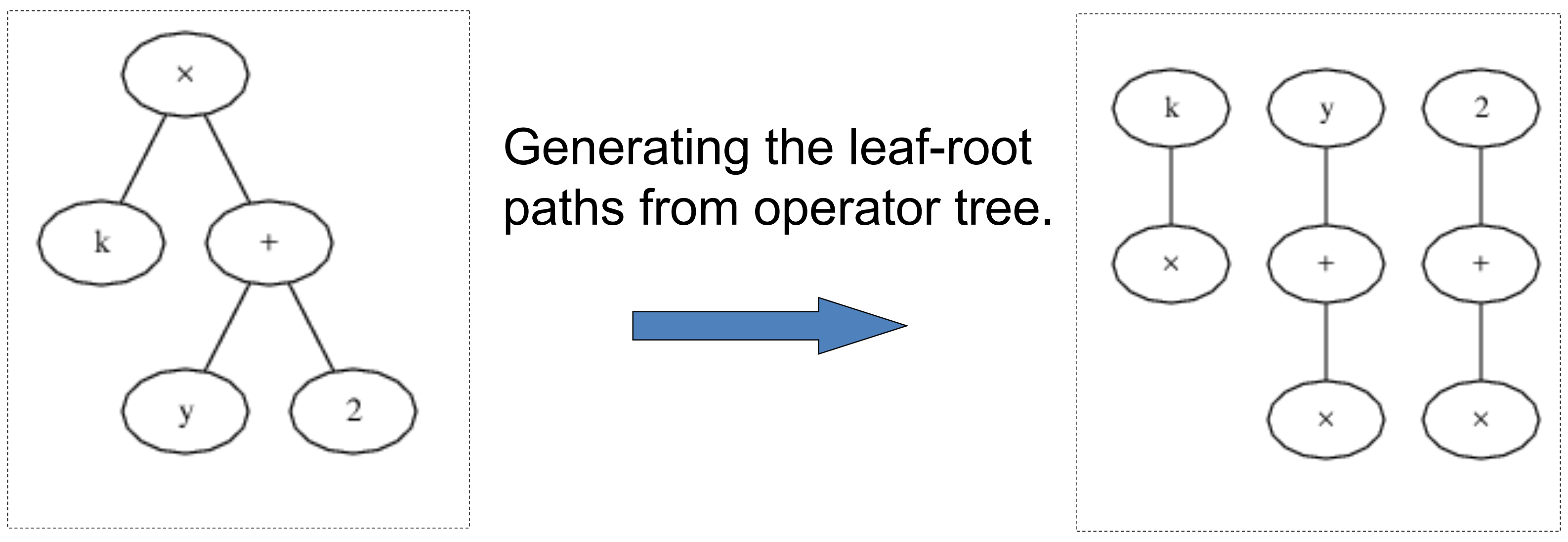
step 1: Crawl and extract math expressions in LaTeX



step 2: Convert LaTeX math expressions into operator tree using a LALR parser.

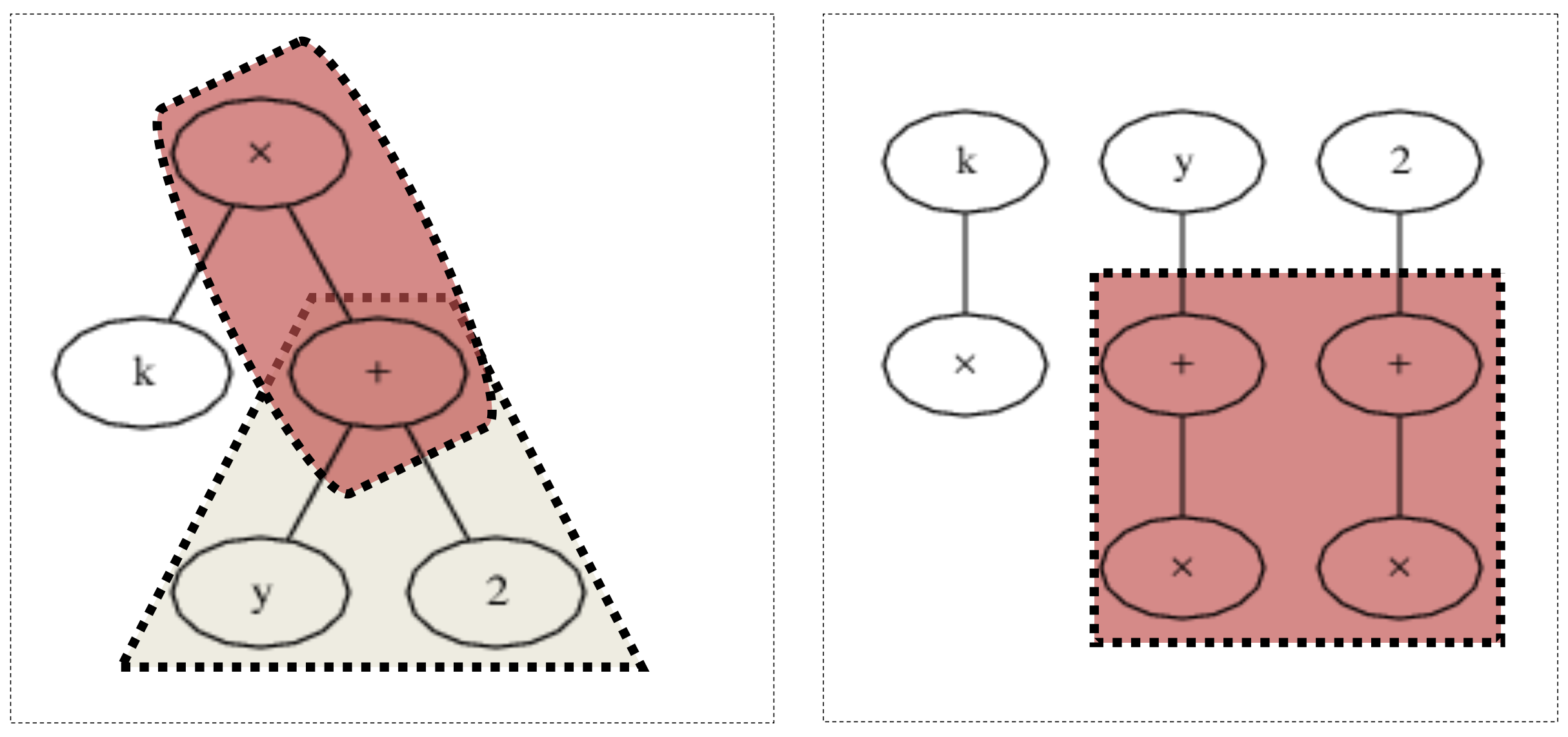
Leaf-root Path and Subtree Properties

The leaf-root path from operator tree is heavily used in our system because an operator tree uniquely determines the leaf-root paths decomposed from the tree, no matter how the operands are ordered.



Operator tree of $k\times(y+2)$ Generated paths

Leaf-root Path and Subtree Properties (Cont.)

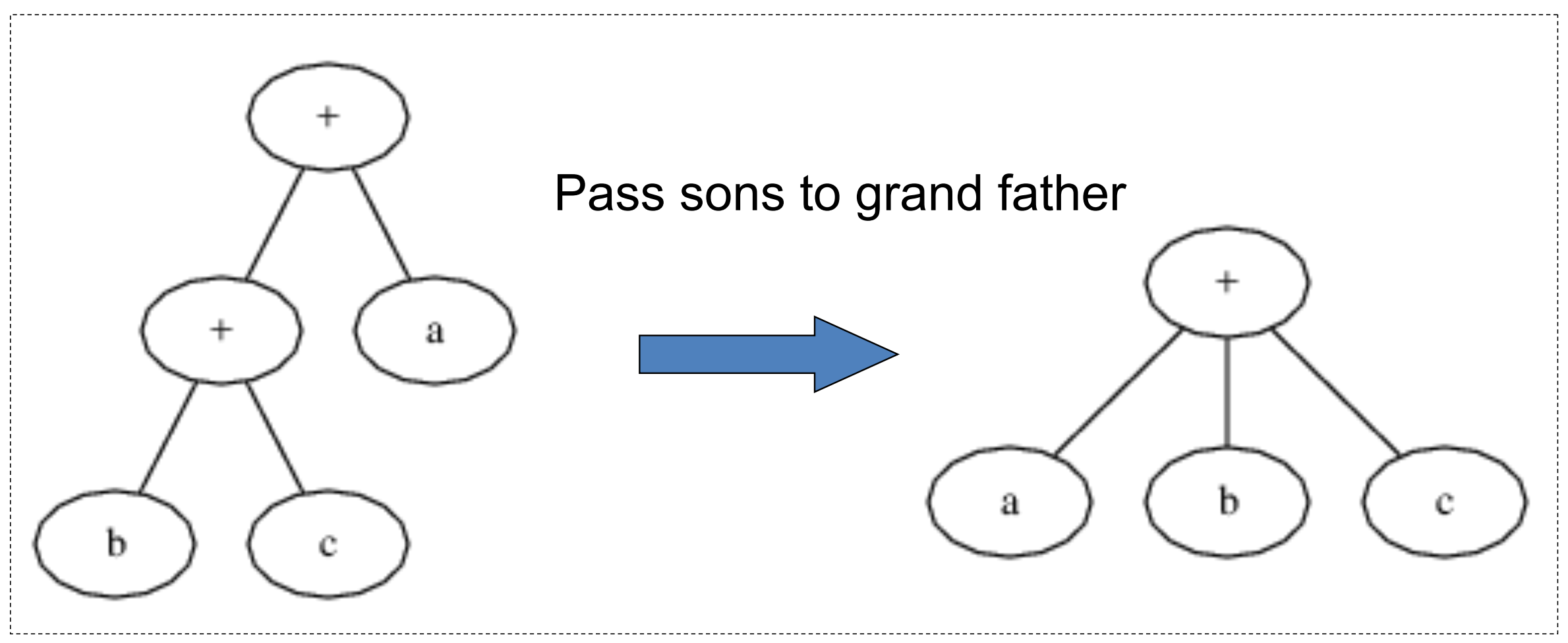


Operator tree Generated paths

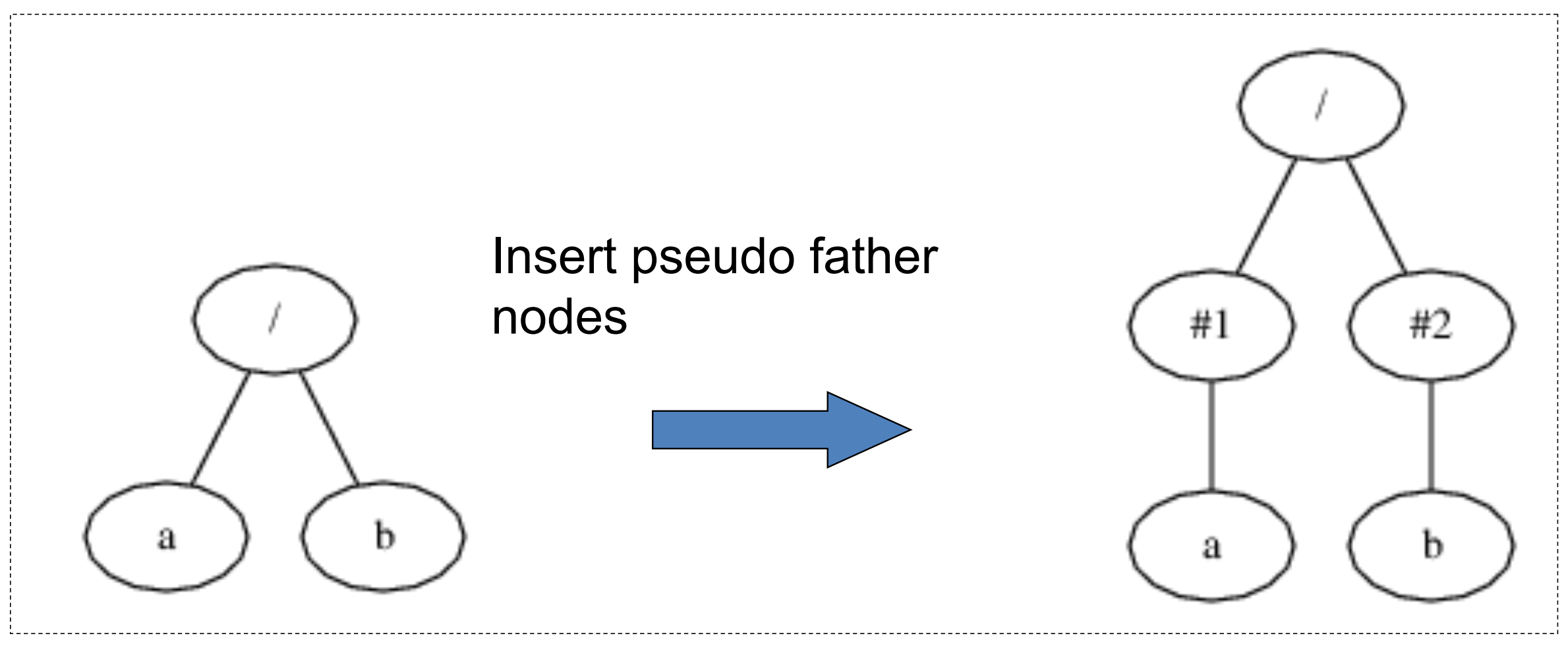
Notice leaf-root paths from the same subtree must share some common nodes (from the root of parent tree to the root of subtree).

We further ensure the subtree of an operator tree T also represents the sub-expression of the expression which T represents:

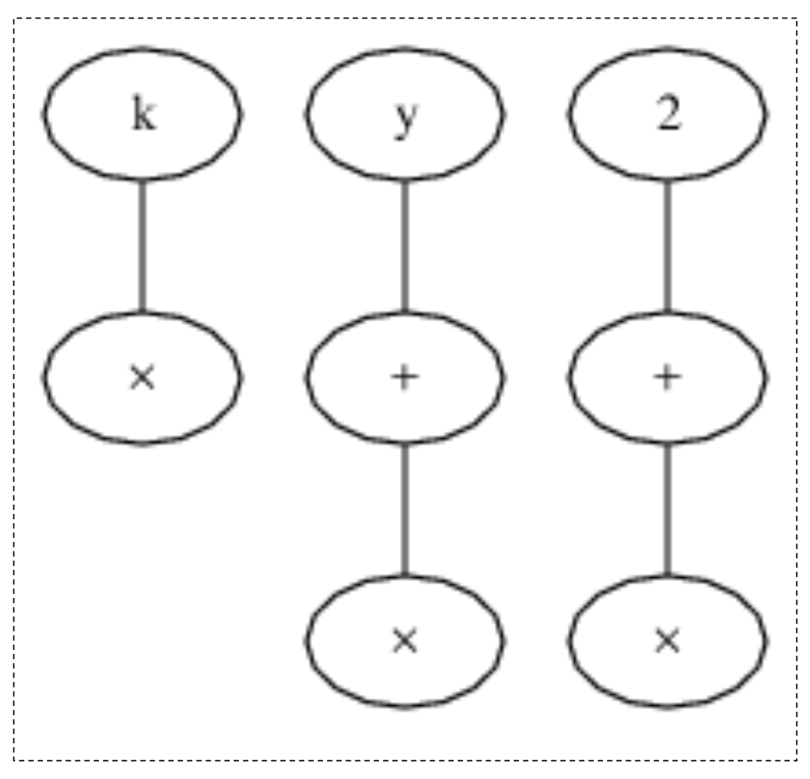
Case 1: If a commutative node has a father operator who is also commutative, the node will pass its children to its father and delete itself.



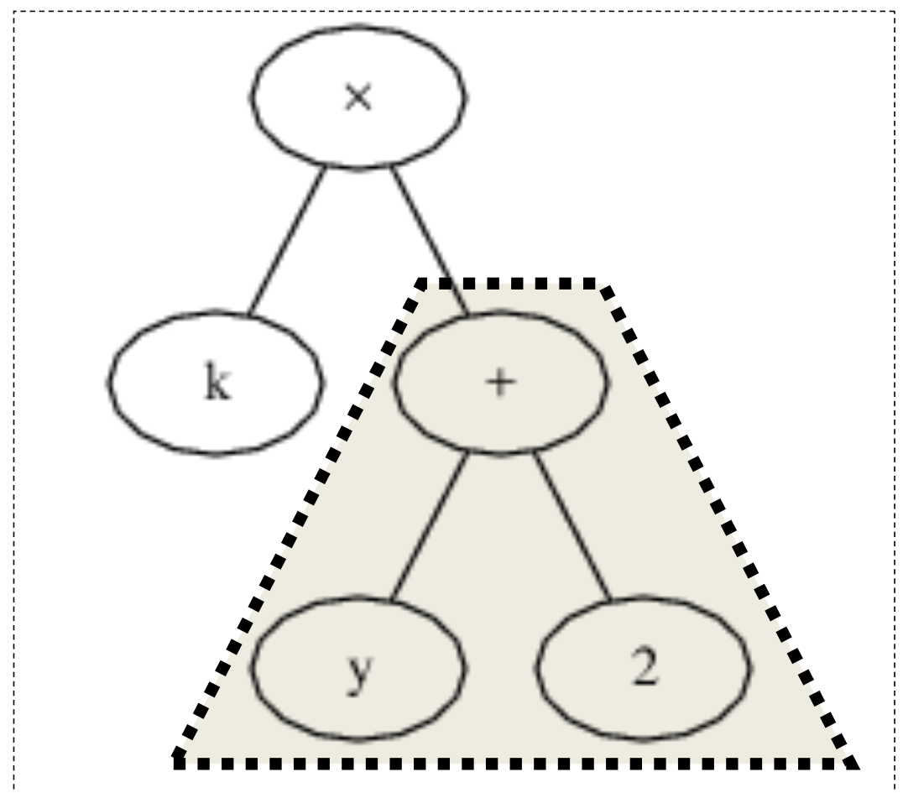
Case 2: When non-commutative operator is being constructed, insert pseudo nodes on top of its children.



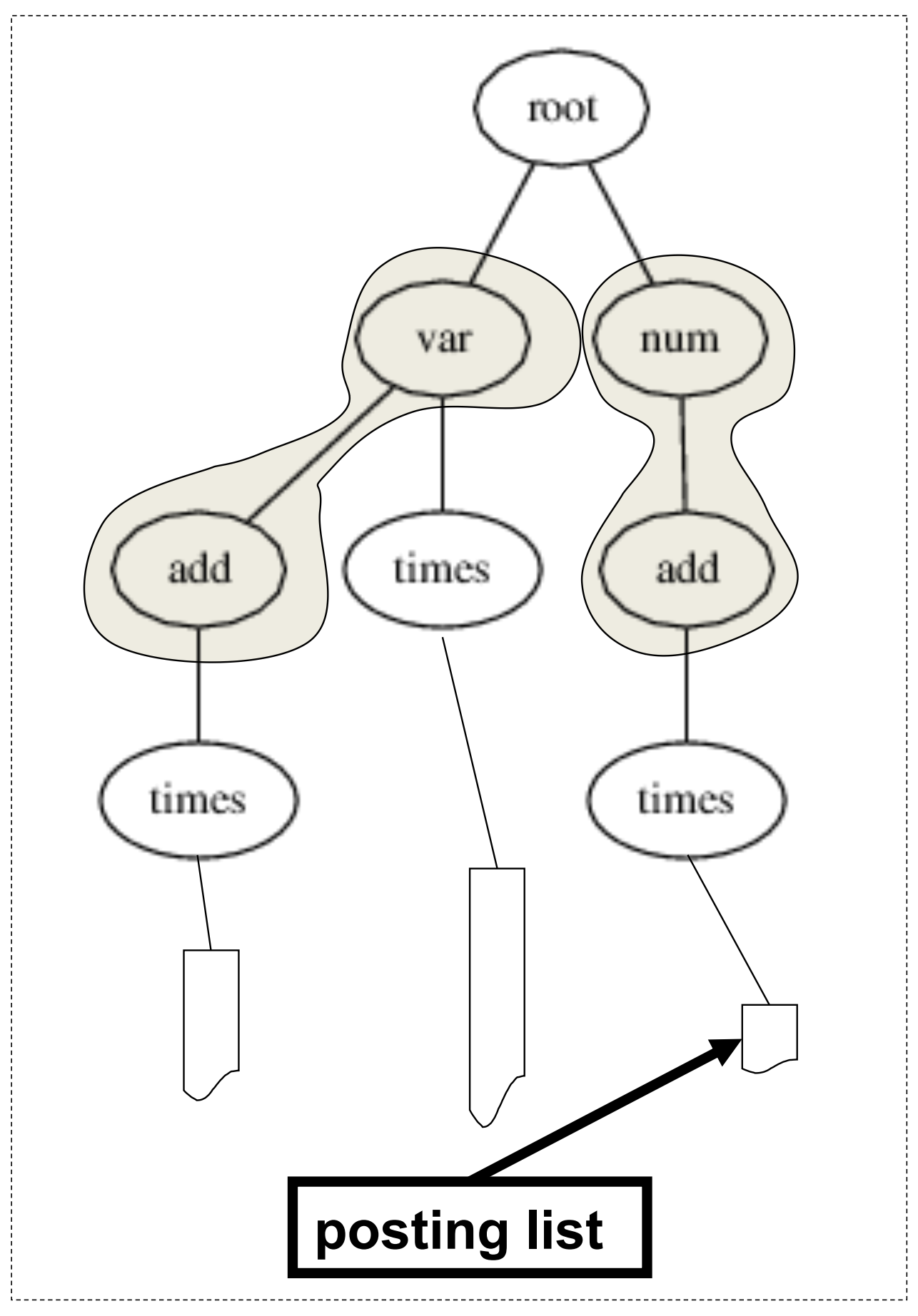
Index and Search



Index: Tokenize leaf-root paths and insert into a persistent on-disk tree-structured index.

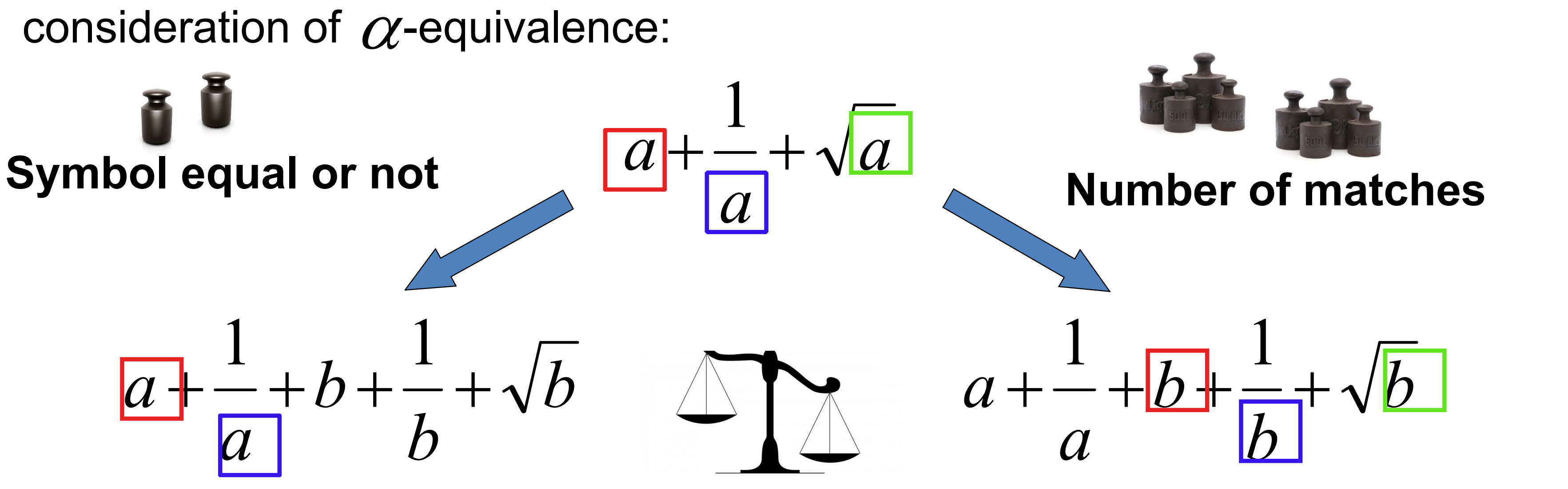


Search: Query a sub-expression and search recursively (merge along the way to the posing lists) in the on-disk index tree.



Ranking Score

Mark-and-Cross algorithm [1] is used to score relevance degree in terms of symbol set similarity between document and query expression, also with the consideration of α-equivalence:


Symbol equal or not
$\boxed{a}+\frac{1}{\boxed{a}}+\sqrt{\boxed{a}}$
Number of matches
$\boxed{a}+\frac{1}{\boxed{a}}+b+\frac{1}{b}+\sqrt{b}$
$a+\frac{1}{a}+\boxed{b}+\frac{1}{\boxed{b}}+\sqrt{\boxed{b}}$

To our perception, more structural matches implies more score, symbol match in math is relatively less important.

References

[1] Wei Zhong. A Novel Similarity-Search Method for Mathematical Content in LaTeX Markup and Its Implementation. http://tkhost.github.io/opmes/thesis-ref.pdf, 2015.