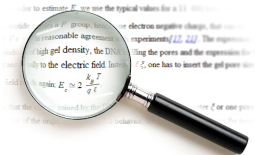


# Math Indexer and Searcher under the Hood: Fine-tuning Query Expansion and Unification Strategies

Michal Růžička, Petr Sojka, Martin Liška

Masaryk University, Faculty of Informatics, Brno, Czech Republic  
mruczicka@mail.muni.cz, sojka@fi.muni.cz, 255768@mail.muni.cz

<https://mir.fi.muni.cz/>



Illustrations by Jiří Franek.

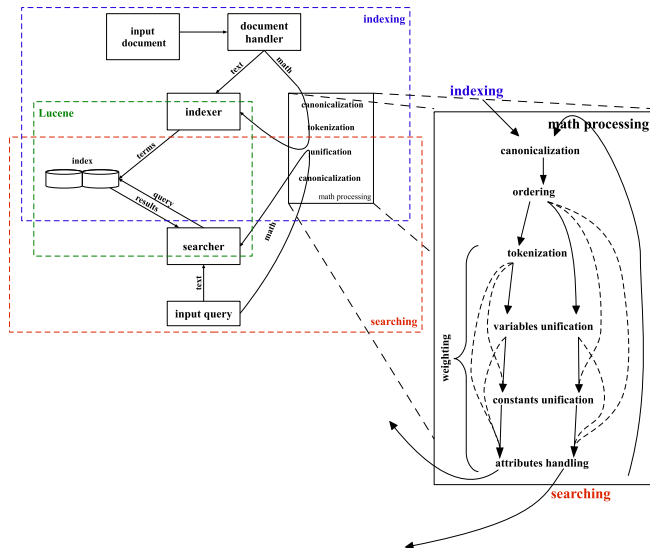
# Outline

- 1 MIaS Overview
- 2 New Approaches tried for NTCIR-12
- 3 Run Selection, Evaluation and Results
- 4 Summary and Future Research Directions

# Outline

- 1 MlaS Overview
- 2 New Approaches tried for NTCIR-12
- 3 Run Selection, Evaluation and Results
- 4 Summary and Future Research Directions

# MIA S Overview



# MIA S Formulae Representation and Weighting

**input:**

$$(a + b^{2+c}, 0.125)$$

**ordering:**

$$(a + b^{c+2}, 0.125)$$

**tokenization:**

$$(a, 0.0875)$$

$$(+, 0.0875)$$

$$(b^{c+2}, 0.0875)$$

$$(b, 0.06125)$$

$$(c+2, 0.06125)$$

$$(c, 0.042875)$$

$$(+, 0.042875)$$

$$(2, 0.042875)$$

**variables unification:**

$$(id_1 + id_2^{id_3+2}, 0.1)$$

$$(id_1^{id_3+2}, 0.07)$$

$$(id_1 + 2, 0.0343)$$

**constants unification:**

$$(a + b^{c+const}, 0.0625)$$

$$(b^{c+const}, 0.04375)$$

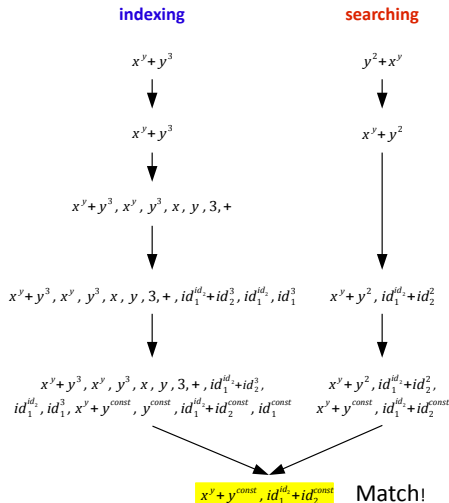
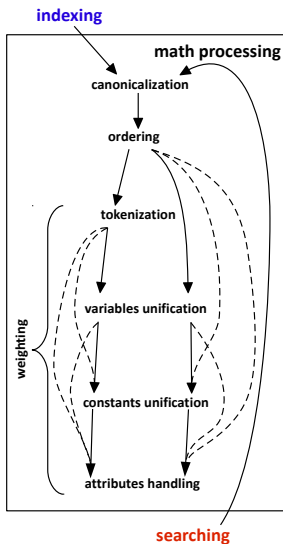
$$(c + const, 0.030625)$$

$$(id_1 + id_2^{id_3+const}, 0.05)$$

$$(id_1^{id_3+const}, 0.035)$$

$$(id_1 + const, 0.01715)$$

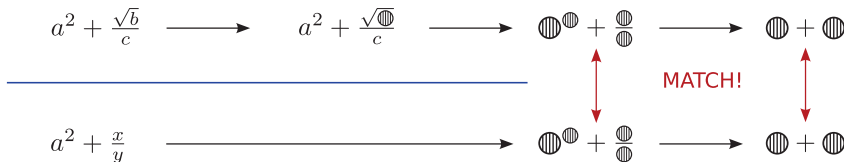
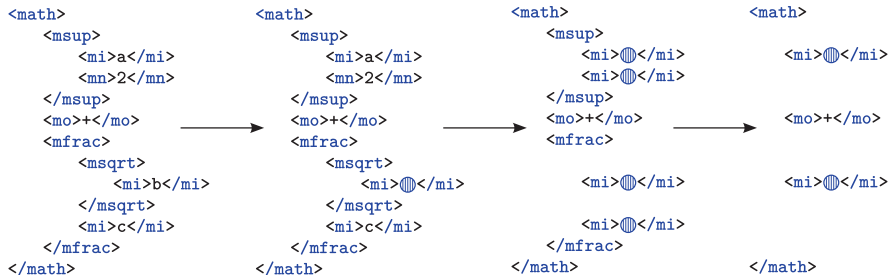
# MIA S Formulae Searching and Matching



# Outline

- 1 MlaS Overview
- 2 New Approaches tried for NTCIR-12**
- 3 Run Selection, Evaluation and Results
- 4 Summary and Future Research Directions

# Structural Unification





## Query Relaxation: Deriving Multiple Queries from the Original

Example of query consisting of two formulae and three text keywords:

$$f_1: s(p) = id$$

$$f_2: p \neq id$$

$k_1$ : "hyperelliptic surface"

$k_2$ : "Riemann"

$k_3$ : "holomorphic map"

original query:  $(f_1 \vee f_2) \wedge (k_1 \vee k_2 \vee k_3)$

subquery 1:  $(f_1 \vee f_2) \wedge (k_1 \vee k_2)$

subquery 2:  $(f_1 \vee f_2) \wedge (k_1)$

subquery 3:  $(f_1 \vee f_2)$

subquery 4:  $(f_1) \wedge (k_1 \vee k_2 \vee k_3)$

subquery 5:  $(k_1 \vee k_2 \vee k_3)$

Notice the strategy is not symmetric.

## Query Relaxation: Deriving Multiple Queries from the Original

Example of query consisting of two formulae and three text keywords:

$$f_1: s(p) = id$$

$$f_2: p \neq id$$

$k_1$ : "hyperelliptic surface"

$k_2$ : "Riemann"

$k_3$ : "holomorphic map"

original query:  $(f_1 \vee f_2) \wedge (k_1 \vee k_2 \vee k_3)$

subquery 1:  $(f_1 \vee f_2) \wedge (k_1 \vee k_2)$

subquery 2:  $(f_1 \vee f_2) \wedge (k_1)$

subquery 3:  $(f_1 \vee f_2)$

subquery 4:  $(f_1) \wedge (k_1 \vee k_2 \vee k_3)$

subquery 5:  $(k_1 \vee k_2 \vee k_3)$

Notice the strategy is not symmetric.

## Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$
- 6:  $r^6_{\text{subquery 2}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{subquery 1}}$
- 11:  $r^{11}_{\text{subquery 1}}$
- 12:  $r^4_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^7_{\text{original}}$
- 16:  $r^8_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^4_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$
- 6:  $r^6_{\text{subquery 2}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{subquery 1}}$
- 11:  $r^{11}_{\text{subquery 1}}$
- 12:  $r^4_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^7_{\text{original}}$
- 16:  $r^8_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^4_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$
- 6:  $r^6_{\text{subquery 2}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{subquery 1}}$
- 11:  $r^{11}_{\text{subquery 1}}$
- 12:  $r^4_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^7_{\text{original}}$
- 16:  $r^8_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^4_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$
- 6:  $r^6_{\text{subquery 2}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{subquery 1}}$
- 11:  $r^{11}_{\text{subquery 1}}$
- 12:  $r^4_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^7_{\text{original}}$
- 16:  $r^8_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^4_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^1_{\text{subquery 1}}$
- 5:  $r^2_{\text{subquery 1}}$
- 6:  $r^1_{\text{subquery 2}}$
- 7:  $r^4_{\text{original}}$
- 8:  $r^5_{\text{original}}$
- 9:  $r^6_{\text{original}}$
- 10:  $r^3_{\text{subquery 1}}$
- 11:  $r^4_{\text{subquery 1}}$
- 12:  $r^2_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^9_{\text{original}}$
- 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^1_{\text{subquery 1}}$
- 5:  $r^2_{\text{subquery 1}}$
- 6:  $r^1_{\text{subquery 2}}$
- 7:  $r^4_{\text{original}}$
- 8:  $r^5_{\text{original}}$
- 9:  $r^6_{\text{original}}$
- 10:  $r^3_{\text{subquery 1}}$
- 11:  $r^4_{\text{subquery 1}}$
- 12:  $r^2_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^9_{\text{original}}$
- 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.



# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^1_{\text{subquery 1}}$
- 5:  $r^2_{\text{subquery 1}}$
- 6:  $r^1_{\text{subquery 2}}$
- 7:  $r^4_{\text{original}}$
- 8:  $r^5_{\text{original}}$
- 9:  $r^6_{\text{original}}$
- 10:  $r^3_{\text{subquery 1}}$
- 11:  $r^4_{\text{subquery 1}}$
- 12:  $r^2_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^9_{\text{original}}$
- 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^1_{\text{subquery 1}}$
- 5:  $r^2_{\text{subquery 1}}$
- 6:  $r^1_{\text{subquery 2}}$
- 7:  $r^4_{\text{original}}$
- 8:  $r^5_{\text{original}}$
- 9:  $r^6_{\text{original}}$
- 10:  $r^3_{\text{subquery 1}}$
- 11:  $r^4_{\text{subquery 1}}$
- 12:  $r^2_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^9_{\text{original}}$
- 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^1_{\text{subquery 1}}$
- 5:  $r^2_{\text{subquery 1}}$
- 6:  $r^1_{\text{subquery 2}}$
- 7:  $r^4_{\text{original}}$
- 8:  $r^5_{\text{original}}$
- 9:  $r^6_{\text{original}}$
- 10:  $r^3_{\text{subquery 1}}$
- 11:  $r^4_{\text{subquery 1}}$
- 12:  $r^2_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^9_{\text{original}}$
- 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r_1^{\text{original}}$
- 2:  $r_2^{\text{original}}$
- 3:  $r_3^{\text{original}}$
- 4:  $r_4^{\text{original}}$
- 5:  $r_5^{\text{original}}$
- 6:  $r_6^{\text{original}}$
- 7:  $r_7^{\text{original}}$
- 8:  $r_8^{\text{original}}$
- 9:  $r_9^{\text{original}}$
- 10:  $r_{10}^{\text{original}}$
- 11:  $r_{11}^{\text{original}}$

*Results of the subquery 1:*

- 1:  $r_1^{\text{subquery 1}}$
- 2:  $r_2^{\text{subquery 1}}$
- 3:  $r_3^{\text{subquery 1}}$
- 4:  $r_4^{\text{subquery 1}}$
- 5:  $r_5^{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r_1^{\text{subquery 2}}$
- 2:  $r_2^{\text{subquery 2}}$
- 3:  $r_3^{\text{subquery 2}}$
- 4:  $r_4^{\text{subquery 2}}$
- 5:  $r_5^{\text{subquery 2}}$

*The final result list:*

- 1:  $r_1^{\text{original}}$
- 2:  $r_2^{\text{original}}$
- 3:  $r_3^{\text{original}}$
- 4:  $r_1^{\text{subquery 1}}$
- 5:  $r_2^{\text{subquery 1}}$
- 6:  $r_1^{\text{subquery 2}}$
- 7:  $r_4^{\text{original}}$
- 8:  $r_5^{\text{original}}$
- 9:  $r_6^{\text{original}}$
- 10:  $r_3^{\text{subquery 1}}$
- 11:  $r_4^{\text{subquery 1}}$
- 12:  $r_2^{\text{subquery 2}}$
- 13:  $r_7^{\text{original}}$
- 14:  $r_8^{\text{original}}$
- 15:  $r_9^{\text{original}}$
- 16:  $r_5^{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r_3^{\text{subquery 2}}$
- 18:  $r_{10}^{\text{original}}$
- 19:  $r_{11}^{\text{original}}$
- No more results from the original query.
- 20:  $r_4^{\text{subquery 2}}$
- 21:  $r_5^{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^1_{\text{subquery 1}}$
- 5:  $r^2_{\text{subquery 1}}$
- 6:  $r^1_{\text{subquery 2}}$
- 7:  $r^4_{\text{original}}$
- 8:  $r^5_{\text{original}}$
- 9:  $r^6_{\text{original}}$
- 10:  $r^3_{\text{subquery 1}}$
- 11:  $r^4_{\text{subquery 1}}$
- 12:  $r^2_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^9_{\text{original}}$
- 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^1_{\text{subquery 1}}$
- 5:  $r^2_{\text{subquery 1}}$
- 6:  $r^1_{\text{subquery 2}}$
- 7:  $r^4_{\text{original}}$
- 8:  $r^5_{\text{original}}$
- 9:  $r^6_{\text{original}}$
- 10:  $r^3_{\text{subquery 1}}$
- 11:  $r^4_{\text{subquery 1}}$
- 12:  $r^2_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^9_{\text{original}}$
- 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
  - 2:  $r^2_{\text{original}}$
  - 3:  $r^3_{\text{original}}$
  - 4:  $r^1_{\text{subquery 1}}$
  - 5:  $r^2_{\text{subquery 1}}$
  - 6:  $r^1_{\text{subquery 2}}$
  - 7:  $r^4_{\text{original}}$
  - 8:  $r^5_{\text{original}}$
  - 9:  $r^6_{\text{original}}$
  - 10:  $r^3_{\text{subquery 1}}$
  - 11:  $r^4_{\text{subquery 1}}$
  - 12:  $r^2_{\text{subquery 2}}$
  - 13:  $r^7_{\text{original}}$
  - 14:  $r^8_{\text{original}}$
  - 15:  $r^9_{\text{original}}$
  - 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
  - 18:  $r^{10}_{\text{original}}$
  - 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
  - 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^1_{\text{subquery 1}}$
- 5:  $r^2_{\text{subquery 1}}$
- 6:  $r^1_{\text{subquery 2}}$
- 7:  $r^4_{\text{original}}$
- 8:  $r^5_{\text{original}}$
- 9:  $r^6_{\text{original}}$
- 10:  $r^3_{\text{subquery 1}}$
- 11:  $r^4_{\text{subquery 1}}$
- 12:  $r^2_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^9_{\text{original}}$
- 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.



# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
  - 2:  $r^2_{\text{original}}$
  - 3:  $r^3_{\text{original}}$
  - 4:  $r^1_{\text{subquery 1}}$
  - 5:  $r^2_{\text{subquery 1}}$
  - 6:  $r^1_{\text{subquery 2}}$
  - 7:  $r^4_{\text{original}}$
  - 8:  $r^5_{\text{original}}$
  - 9:  $r^6_{\text{original}}$
  - 10:  $r^3_{\text{subquery 1}}$
  - 11:  $r^4_{\text{subquery 1}}$
  - 12:  $r^2_{\text{subquery 2}}$
  - 13:  $r^7_{\text{original}}$
  - 14:  $r^8_{\text{original}}$
  - 15:  $r^9_{\text{original}}$
  - 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
  - 18:  $r^{10}_{\text{original}}$
  - 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
  - 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^1_{\text{subquery 1}}$
- 5:  $r^2_{\text{subquery 1}}$
- 6:  $r^1_{\text{subquery 2}}$
- 7:  $r^4_{\text{original}}$
- 8:  $r^5_{\text{original}}$
- 9:  $r^6_{\text{original}}$
- 10:  $r^3_{\text{subquery 1}}$
- 11:  $r^4_{\text{subquery 1}}$
- 12:  $r^2_{\text{subquery 2}}$
- 13:  $r^7_{\text{original}}$
- 14:  $r^8_{\text{original}}$
- 15:  $r^9_{\text{original}}$
- 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
- 18:  $r^{10}_{\text{original}}$
- 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
- 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

# Strip Merging of Search Engine Results Pages (SERPs)

original query:  $(f_1) \wedge (k_1)$   
 subquery 1:  $(f_1)$   
 subquery 2:  $(k_1)$

*Results of the original query:*

- 1:  $r^1_{\text{original}}$
- 2:  $r^2_{\text{original}}$
- 3:  $r^3_{\text{original}}$
- 4:  $r^4_{\text{original}}$
- 5:  $r^5_{\text{original}}$
- 6:  $r^6_{\text{original}}$
- 7:  $r^7_{\text{original}}$
- 8:  $r^8_{\text{original}}$
- 9:  $r^9_{\text{original}}$
- 10:  $r^{10}_{\text{original}}$
- 11:  $r^{11}_{\text{original}}$

*Results of the subquery 1:*

- 1:  $r^1_{\text{subquery 1}}$
- 2:  $r^2_{\text{subquery 1}}$
- 3:  $r^3_{\text{subquery 1}}$
- 4:  $r^4_{\text{subquery 1}}$
- 5:  $r^5_{\text{subquery 1}}$

*Results of the subquery 2:*

- 1:  $r^1_{\text{subquery 2}}$
- 2:  $r^2_{\text{subquery 2}}$
- 3:  $r^3_{\text{subquery 2}}$
- 4:  $r^4_{\text{subquery 2}}$
- 5:  $r^5_{\text{subquery 2}}$

*The final result list:*

- 1:  $r^1_{\text{original}}$
  - 2:  $r^2_{\text{original}}$
  - 3:  $r^3_{\text{original}}$
  - 4:  $r^1_{\text{subquery 1}}$
  - 5:  $r^2_{\text{subquery 1}}$
  - 6:  $r^1_{\text{subquery 2}}$
  - 7:  $r^4_{\text{original}}$
  - 8:  $r^5_{\text{original}}$
  - 9:  $r^6_{\text{original}}$
  - 10:  $r^3_{\text{subquery 1}}$
  - 11:  $r^4_{\text{subquery 1}}$
  - 12:  $r^2_{\text{subquery 2}}$
  - 13:  $r^7_{\text{original}}$
  - 14:  $r^8_{\text{original}}$
  - 15:  $r^9_{\text{original}}$
  - 16:  $r^5_{\text{subquery 1}}$
- No more results from subquery 1.
- 17:  $r^3_{\text{subquery 2}}$
  - 18:  $r^{10}_{\text{original}}$
  - 19:  $r^{11}_{\text{original}}$
- No more results from the original query.
- 20:  $r^4_{\text{subquery 2}}$
  - 21:  $r^5_{\text{subquery 2}}$
- No more results from subquery 2.

## New Querying Strategies and Result Merging

### NTCIR-11 Reverse (*N11Rev*)

original:	$f_1$	$k_1$	$k_2$	$k_3$
1:	$f_1$		$k_2$	$k_3$
2:	$f_1$			$k_3$
3:	$f_1$			
4:		$k_1$	$k_2$	$k_3$

### NTCIR-11 Remove Most Specific First (*N11RmSpecFrst*)

original:	$f_1$	$k_1$	$k_2$	$k_3$
1:	$f_1$		$k_2$	$k_3$
2:	$f_1$		$k_2$	
3:	$f_1$			
4:		$k_1$	$k_2$	$k_3$

### NTCIR-11 Random Input Reordering (*N11Rnd*)

original:	$f_1$	$k_1$	$k_2$	$k_3$
1:	$f_1$	$k_1$	$k_2$	
2:	$f_1$		$k_2$	
3:	$f_1$			
4:		$k_1$	$k_2$	$k_3$

### NTCIR-11 Remove Most Specific Last (*N11RmSpecLast*)

original:	$f_1$	$k_1$	$k_2$	$k_3$
1:	$f_1$	$k_1$		$k_3$
2:	$f_1$	$k_1$		
3:	$f_1$			
4:		$k_1$	$k_2$	$k_3$

Colours denote the specificity of the term: term resulting in **high number of results** is **less specific** than **highly specific term** resulting in **medium** or **low number of results**.

# New Querying Strategies and Result Merging

## Original Query Only (*OrigOnly*)

 $f_1 \quad k_1 \quad k_2 \quad k_3$ 

## Math Terms Only (*OrigOnlyMath*)

 $f_1$ 

## Text Terms Only (*OrigOnlyText*)

 $k_1 \quad k_2 \quad k_3$ 

## All Possible Subqueries (*AllPsbQrs*)

original:	$f_1$	$k_1$	$k_2$	$k_3$
1:	$f_1$	$k_1$	$k_2$	
2:	$f_1$		$k_2$	$k_3$
3:	$f_1$	$k_1$		$k_3$
4:	$f_1$	$k_1$		
5:		$k_1$	$k_2$	$k_3$
6:	$f_1$		$k_2$	
7:	$f_1$			$k_3$
8:			$k_2$	$k_3$
9:		$k_1$		$k_3$
10:		$k_1$	$k_2$	
11:	$f_1$			
12:		$k_1$		
13:			$k_2$	
14:				$k_3$

# New Querying Strategies and Result Merging

## Leave One Out (*OneOut*)

original:	$f_1$	$k_1$	$k_2$	$k_3$
1:	$f_1$	$k_1$	$k_2$	
2:	$f_1$		$k_2$	$k_3$
3:	$f_1$	$k_1$		$k_3$
4:		$k_1$	$k_2$	$k_3$

## Leave One or Two Out (*OneTwoOut*)

original:	$f_1$	$k_1$	$k_2$	$k_3$
1:	$f_1$		$k_2$	$k_3$
2:	$f_1$	$k_1$		$k_3$
3:	$f_1$	$k_1$	$k_2$	
4:	$f_1$			$k_3$
5:	$f_1$		$k_2$	
6:		$k_1$	$k_2$	$k_3$
7:	$f_1$	$k_1$		
8:		$k_1$		$k_3$
9:		$k_1$	$k_2$	



## Phrase Expansion

These modifications are only applicable on multi-word text keywords of the original query. For example, having query

**Formula 1:**  $\aleph_0$

**Keyword 1:** categorical simple theory

we have two methods of transformation — *phrase* and *full phrase expansion*.

# Phrase Expansion

**Phrase expansion** For multi-word keywords individual words are used instead of the original multi-word keywords.

Thus, our query is transformed to query:

**Formula 1:**  $\aleph_0$

**Keyword 1:** categorical

**Keyword 2:** simple

**Keyword 3:** theory



# Phrase Expansion

**Full phrase expansion** Individual words from the multi-word keywords are added one by one at the end of the keywords list (removing duplicates across multiple multi-keywords, if any).

Thus, our query is transformed to query:

**Formula 1:**  $\aleph_0$

**Keyword 1:** categorical simple theory

**Keyword 2:** categorical

**Keyword 3:** simple

**Keyword 4:** theory

# Math Handling Features

**Canonicalization** The canonicalization process aims to normalize potential serializations (different notations in MathML encoding) of the same math formulae. The normalization is optimized for similarity search not to preserve full semantic information of the original formulae but possibly removes semantically negligible differences in behalf of similarity matches.

# Math Handling Features

**Canonicalization operators removal** List of math operators to be removed from canonicalized formulae:

- U+2062 INVISIBLE TIMES
- U+22C5 DOT OPERATOR
- U+002A ASTERISK
- U+2063 INVISIBLE SEPARATOR
- U+2064 INVISIBLE PLUS

## Math Handling Features

- Unary operators removal** Unary operators are removed from the input formulae in the process of formulae normalization by our MathML Canonicalizer.
- Operator unification** We define an operator equivalence relation. We substitute all of these operators (e.g.  $+$ ,  $-$ ) with a canonical operator that represents each equivalence class.
- Structural unification** Indexing structurally unified derivatives of the original formulae was used for the first time by MiaS system at NTCIR-12.

# Outline

- 1 MlaS Overview
- 2 New Approaches tried for NTCIR-12
- 3 Run Selection, Evaluation and Results**
- 4 Summary and Future Research Directions

# NTCIR-11 Evaluation and NTCIR-12 Runs Selection

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	BPREF	PCM-T	CM-APSWMWWS	CM-LOOTO	CM-LOO	CM-N1MRIR	CM-N1MRMSFR	CM-N1MRMSF	CM-N1MR	CM-N1M	CM-OQFPO	CM-OQKPO	CM-OQO
4	ntcir-12-10-pfe	0,0492	0,1029	0,1011	0,0987	0,1047	0,1047	0,1057	0,1027	0,1055	0,0220	0,0619	0,0343
5	ntcir-12-11	0,0557	0,0897	0,0873	0,0898	0,0972	0,0977	0,0973	0,0956	0,0977	0,0216	0,0514	0,0348
6	ntcir-12-11-fe	0,0529	0,0964	0,0991	0,1040	0,1105	0,1104	0,1107	0,1079	0,1110	0,0216	0,0673	0,0341
7	ntcir-12-11-pfe	0,0509	0,1037	0,1015	0,0987	0,1045	0,1038	0,1055	0,1019	0,1051	0,0216	0,0616	0,0348
8	ntcir-12-12	0,0552	0,0893	0,0866	0,0891	0,0970	0,0971	0,0969	0,0951	0,0969	0,0226	0,0511	0,0367
9	ntcir-12-12-fe	0,0527	0,0949	0,0969	0,1023	0,1094	0,1090	0,1097	0,1073	0,1094	0,0226	0,0670	0,0351
10	ntcir-12-12-pfe	0,0502	0,1009	0,0989	0,0976	0,1040	0,1033	0,1045	0,1015	0,1040	0,0226	0,0614	0,0357
11	ntcir-12-13	0,0561	0,0882	0,0860	0,0889	0,0960	0,0967	0,0962	0,0945	0,0970	0,0224	0,0513	0,0373
12	ntcir-12-13-fe	0,0529	0,0963	0,0981	0,1032	0,1089	0,1096	0,1094	0,1067	0,1099	0,0224	0,0671	0,0363
13	ntcir-12-13-pfe	0,0517	0,1014	0,1012	0,0986	0,1035	0,1034	0,1043	0,1013	0,1045	0,0224	0,0616	0,0368
14	ntcir-12-14	0,0548	0,0871	0,0858	0,0872	0,0944	0,0945	0,0947	0,0927	0,0947	0,0221	0,0511	0,0356
15	ntcir-12-14-fe	0,0517	0,0935	0,0954	0,1008	0,1066	0,1071	0,1078	0,1046	0,1075	0,0221	0,0668	0,0340
16	ntcir-12-14-pfe	0,0503	0,0995	0,0988	0,0965	0,1021	0,1014	0,1029	0,0996	0,1024	0,0221	0,0613	0,0350
17	ntcir-12-15	0,0965	0,0892	0,0870	0,0895	0,0984	0,0986	0,0984	0,0966	0,0987	0,0220	0,0515	0,0340
18	ntcir-12-15-fe	0,1067	0,0942	0,0980	0,1033	0,1096	0,1102	0,1102	0,1081	0,1106	0,0220	0,0671	0,0332
19	ntcir-12-15-pfe	0,1021	0,1013	0,1001	0,0979	0,1044	0,1040	0,1052	0,1020	0,1048	0,0220	0,0614	0,0338
20	ntcir-12-16	0,0955	0,0884	0,0870	0,0895	0,0974	0,0976	0,0980	0,0961	0,0974	0,0225	0,0512	0,0347
21	ntcir-12-16-fe	0,1058	0,0961	0,0988	0,1038	0,1097	0,1100	0,1104	0,1079	0,1104	0,0225	0,0668	0,0340
22	ntcir-12-16-pfe	0,1017	0,1032	0,1013	0,0989	0,1042	0,1040	0,1051	0,1019	0,1046	0,0225	0,0614	0,0345
23	ntcir-12-17	0,0665	0,0662	0,0682	0,0677	0,0653	0,0639	0,0671	0,0642	0,0659	0,0085	0,0512	0,0411
24	ntcir-12-17-fe	0,0579	0,0602	0,0646	0,0661	0,0616	0,0594	0,0597	0,0578	0,0590	0,0085	0,0670	0,0388
25	ntcir-12-17-pfe	0,0570	0,0685	0,0708	0,0706	0,0653	0,0654	0,0662	0,0640	0,0674	0,0085	0,0615	0,0440
26	ntcir-12-18	0,0467	0,0438	0,0443	0,0446	0,0433	0,0419	0,0434	0,0420	0,0427	0,0017	0,0512	0,0236
27	ntcir-12-18-fe	0,0495	0,0484	0,0509	0,0529	0,0509	0,0484	0,0479	0,0463	0,0481	0,0017	0,0668	0,0285
28	ntcir-12-18-pfe	0,0496	0,0496	0,0540	0,0540	0,0477	0,0459	0,0470	0,0458	0,0466	0,0017	0,0613	0,0273
29	ntcir-12-19	0,0000	0,0697	0,0699	0,0681	0,0688	0,0681	0,0702	0,0679	0,0696	0,0089	0,0514	0,0361
30	ntcir-12-19-fe	0,0000	0,0763	0,0781	0,0794	0,0800	0,0772	0,0771	0,0748	0,0772	0,0089	0,0673	0,0432
31	ntcir-12-19-pfe	0,0000	0,0770	0,0809	0,0794	0,0757	0,0734	0,0745	0,0729	0,0745	0,0089	0,0617	0,0417

## Test Hardware Description

- Physical server (no virtualization).
- 2 × Intel Xeon E5-2650v2 @ 2.60 GHz  
(2 × 8 cores × 2 via hyperthreading).
- 256 GiB of RAM.
- 11 disks 7200 RPM in Btrfs RAID10  
(4 × 6 TB, 4 × 4 TB, 1 × 3 TB, 2 × 1 TB).

# Main Task Indices

Index	Indexing times [min]		Index size [GiB]	Indexed formulae	P@5
	Wall Clock	CPU			
<b>ntcir-12-10</b>	1,054	1,453	47	2,354,850,850	0.1241 (CMATH RMSFR PE)
<b>ntcir-12-15</b>	769	1,155	59	2,704,770,446	–
<b>ntcir-12-18</b>	5,913	5,604	288	5,591,527,950	0.1250 (PCMath LOoTO PE; SimTo)
<b>ntcir-12-19</b>	5,181	5,486	288	5,592,489,129	0.1310 (PCMath LOoTO PE)

Feature	<i>ntcir-12-10</i>	<i>ntcir-12-15</i>	<i>ntcir-12-18</i>	<i>ntcir-12-19</i>
<b>canonicalization</b>	•	–	•	•
<b>special operators removal</b>	•	–	•	•
<b>unary operators removal</b>	–	–	•	•
<b>operator unification</b>	–	–	•	•
<b>structural unification</b>	–	–	•	• (progressive weighting)



# Wiki Task Indices

Index	Indexing times [min]		Index size [GiB]	Indexed formulae	P@5
	Wall Clock	CPU			
<b>ntcir-12-30</b>	30	46	3.4	25,242,248	0.0600 (PMath LOoTO)
<b>ntcir-12-35</b>	20	37	4.7	28,842,062	–
<b>ntcir-12-38</b>	81	98	13	67,231,738	–
<b>ntcir-12-39</b>	82	99	13	67,308,089	0.0600 (PCMath LOoTO PE)

Feature	<i>ntcir-12-30</i>	<i>ntcir-12-35</i>	<i>ntcir-12-38</i>	<i>ntcir-12-39</i>
canonicalization	•	–	•	•
special operators removal	•	–	•	•
unary operators removal	–	–	•	•
operator unification	–	–	•	•
structural unification	–	–	•	• (progressive weighting)

## Effect of Structural Unification

MAP	Number of better results on NTCIR-11 topics	
	With struct. unification	Without struct. unification
no phrase expansion	2	31
phrase expansion	3	30
full phrase expansion	5	25

Bpref@1000	Number of better results on NTCIR-11 topics	
	With struct. unification	Without struct. unification
no phrase expansion	23	10
phrase expansion	24	9
full phrase expansion	14	16

# Outline

- 1 MlaS Overview
- 2 New Approaches tried for NTCIR-12
- 3 Run Selection, Evaluation and Results
- 4 Summary and Future Research Directions**

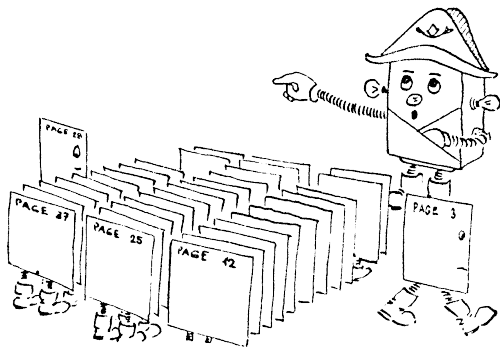
## Summary

- New evaluation platform based on NTCIR-11 ground truth.
- Introduction of math structural unification.
  - Use of structurally unified derivatives increases recall but has negative impact on precision.
  - Fine tuning the weights of structural unification nodes could possibly balance performance of our system towards recall.
  - Setting and tuning the indexing and preprocessing parameters is necessary for given application.

## Future Work and Research Directions

- Machine learning techniques to formulae disambiguation (random walking), formulae weighting, to document ranking and filtering.
- Deploying Computer Algebra Systems for better canonicalization of the input formulae.
- Using NTCIR-11 and NTCIR-12 MathIR ground truth data and our evaluation platform to distinguish effects of MiaS system features.
- Using arXiv.org as a testbed for scalability and development and to provide full text+math search.

# Questions?





Illustrations by Jiří Franek.



RŮŽIČKA, Michal; SOJKA, Petr; LÍŠKA, Martin. Math Indexer and Searcher under the Hood: Fine-Tuning Query Expansion and Unification Strategies. In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2016. 7 pp. [http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/NTCIR/toc\\_ntcir.html](http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings12/NTCIR/toc_ntcir.html)



LÍŠKA, Martin; SOJKA, Petr; RŮŽIČKA, Michal. Combining Text and Formula Queries in Math Information Retrieval: Evaluation of Query Results Merging Strategies. In: Proceedings of the First International Workshop on Novel Web Search Interfaces and Systems. NWSearch'15. Melbourne, Australia: ACM, 2015, p. 7–9. ISBN 978-1-4503-3789-2. DOI: <http://dx.doi.org/10.1145/2810355.2810359>



RŮŽIČKA, Michal; SOJKA, Petr; LÍŠKA, Martin. Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy. In Hidei Joho, Kazuaki Kishida. Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies. Tokyo: National Institute of Informatics, 2014. p. 127–134. ISBN 978-4-86049-065-2. <http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings11/pdf/NTCIR/Math-2/07-NTCIR11-MATH-RuzickaM.pdf>



SOJKA, Petr; LÍŠKA, Martin. The Art of Mathematics Retrieval. In Matthew R. B. Hardy, Frank Wm. Tompa. Proceedings of the 2011 ACM Symposium on Document Engineering. Mountain View, CA, USA: ACM, 2011. p. 57–60, ISBN 978-1-4503-0863-2. DOI: <http://dx.doi.org/10.1145/2034691.2034703>



SOJKA, Petr; LÍŠKA, Martin. Indexing and Searching Mathematics in Digital Libraries – Architecture, Design and Scalability Issues. In James H. Davenport, William M. Farmer, Josef Urban, Florian Rabe. Intelligent Computer Mathematics Lecture Notes in Computer Science, 2011, Volume 6824/2011. Berlin/Heidelberg: Springer, 2011. p. 228–243. ISBN 978-3-642-22672-4. DOI: <http://dx.doi.org/10.1007/978-3-642-22673-116>



LÍŠKA, Martin; SOJKA, Petr; RŮŽIČKA, Michal; MRAVEC, Peter. Web Interface and Collection for Mathematical Retrieval : WebMiaS and MREC. In Petr Sojka, Thierry Bouche. DML 2011: Towards a Digital Mathematics Library. Brno: Masaryk University, 2011. p. 77–84. ISBN 978-80-210-5542-1. <http://dml.cz/handle/10338.dmlcz/702604>