Three is Better than One!
Ensembling Math Information Retrieval Systems

Vít Novotný, Petr Sojka, Michal Štefánik, Dávid Lupták
witiko@mail.muni.cz, sojka@fi.muni.cz

Math Information Retrieval Research Group,
Faculty of Informatics, Masaryk University
https://mir.fi.muni.cz/

September 25, 2020
Task 1: Find Answers
Introduction

For more than a decade now, MIRMU has been grappling with the challenges of MIR:

1. DML-CZ (2008) [1]
2. EuDML (2013) [3]
3. NTCIR (2016) [6, 11, 10]

In ARQMath 2020, we have tackled both task 1 (find answers) and 2 (formula search).

For task 1, we have prepared five MIR systems:

1. Math Indexer and Searcher (MIaS),
2. Soft Cosine Measure (SCM),
3. Formula2Vec,
4. CompuBERT, and
5. Ensemble.
Methods

Math Representations

- In our MIR systems, we used the following math representations:
Methods

Corpora, Relevance Judgements, and Evaluation Measures

For training, we used the following two corpora:


For validation, we used the following two sets of relevance judgements:

1. Automatic (param. opt., model sel.), and 2. Human-Annotated (perf. est.).

In our evaluation, we used the following two measures:

1. Normalized Discounted Cumulative Gain Prime (nDCG'), [12] and 2. Spearman’s Correlation Coefficient ($\rho$).

For retrieval, we used a machine with with 32 CPUs and 252 GiB RAM.

For training embeddings, we used an NVIDIA GTX2080 Ti GPU with 11 GiB VRAM.
Historically the first MIR system deployed in a digital mathematical library. [14]

- Uses TF-IDF with M-Terms extracted from CMML as a math representation.
- **Accuracy:** nDCG’ 0.155, insignificantly below the Tangent-S baseline.
- **Speed:** avg. 1.24 s/topic, min. 0.1 s/topic, max. 7.27 s/topic.
Soft Cosine Measure (SCM)

- Uses TF-IDF with the Prefix math representation and SCM [13, 7, 8] doc. similarity.
- Uses automatic relevance judgements to optimize parameters of fastText and SCM.
- Four different fastText models were trained:
  1. Tiny (5 epochs, alternative submission)
  2. Small (10 epochs, primary submission)
  3. Medium (2 epochs on all corpora)
  4. Large (10 epochs on all corpora)

- **Accuracy:** nDCG’ 0.224 (small), insignificantly below the Approach0 baseline.
- **Speed:** avg. 58.46 s/topic, min. 30.52 s/topic, max. 502.84 s/topic.
**Task 1**  

**Formula2Vec**

- Uses the optimal parameters of fastText and **RedHat defaults** for Doc2Vec.
- Four different Doc2Vec models were trained:
  1. Tiny (5 epochs on no_problem ArXMLiv)
  2. Small (10 epochs, alternative sub.)
  3. Medium (2 epochs on all corpora)
  4. Large (10 epochs on all corpora)

- **Accuracy:** nDCG' 0.050 (small), on par with DPRL and zbMath systems.
- **Speed:** avg. 3.23 s/topic, min. 3.14 s/topic, max. 7.87 s/topic.
Uses sBERT [9] with the \( \LaTeX \) math representation and the cosine similarity.

Uses our automatic relevance judgements to optimize the Triplet objective.

Stark difference in performance between automatic and human-annotated r.j.’s.

**Accuracy:** nDCG’ 0.009, not significantly better than zero.

**Speed:** avg. 3.43 s/topic, min. 3.2 s/topic, max. 3.67 s/topic.
Ensemble

- Interleaves the result lists of primary submissions: MIaS, SCM, and CompuBERT.
- Uses a parameter-free ensembling algorithm that only uses ranks, not scores.
- Results are ranked by median rank, then by frequency, and then interleaved.
- **Tie-breaking:** More than 40% of all results were arbitrarily interleaved.
- **Accuracy:** nDCG' 0.238, best of our systems, significantly better than all but SCM. The ensemble of all non-baseline primary submissions (0.419) best in competition.
**Results**

**Accuracy:** SCM (0.224) significantly better than MIaS, ensemble best in competition.

<table>
<thead>
<tr>
<th></th>
<th>MIA S</th>
<th>SCM</th>
<th>F2Vec</th>
<th>CBRT</th>
<th>Ens.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best¹</td>
<td>0.155</td>
<td>0.237</td>
<td>0.101</td>
<td>0.009</td>
<td>0.419</td>
</tr>
<tr>
<td>Primary</td>
<td>0.155</td>
<td>0.224</td>
<td>0.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alternative</td>
<td></td>
<td></td>
<td>0.050</td>
<td>0.238</td>
<td></td>
</tr>
</tbody>
</table>

**Speed:** On average, MIaS was fastest, SCM slowest, CompuBERT had least variance.

<table>
<thead>
<tr>
<th></th>
<th>MIA S</th>
<th>SCM</th>
<th>F2Vec</th>
<th>CBRT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>0.1</td>
<td>30.52</td>
<td>3.14</td>
<td>3.2</td>
</tr>
<tr>
<td>Average</td>
<td>1.24</td>
<td>58.64</td>
<td>3.23</td>
<td>3.43</td>
</tr>
<tr>
<td>Maximum</td>
<td>7.27</td>
<td>502.84</td>
<td>7.87</td>
<td>3.67</td>
</tr>
</tbody>
</table>

¹Includes unsubmitted out-of-competition results.
Conclusion and Future Work

- We have introduced three significantly different systems:
  1. TF-IDF-based MIaS,
  2. TF-IDF-based SCM, and
  3. CompuBERT.

- TF-IDF-based MIaS and SCM combine high accuracy, speed, [7] and interpretability.
- Transformer-based CompuBERT was highly sensitive to the training objective.
- Three is better than one: ensemble of primary submissions best in competition.
Bibliography


Three is Better than One!
Ensembling Math Information Retrieval Systems

Vít Novotný, Petr Sojka, Michal Štefánik, Dávid Lupták
witiko@mail.muni.cz, sojka@fi.muni.cz

Math Information Retrieval Research Group,
Faculty of Informatics, Masaryk University
https://mir.fi.muni.cz/

September 25, 2020
Task 2: Formula Search
Introduction

For more than a decade now, MIRMU has been grappling with the challenges of MIR:

1. DML-CZ (2008) [1]
2. EuDML (2013) [3]
3. NTCIR (2016) [6, 11, 10]

In ARQMath 2020, we have tackled both task 1 (find answers) and 2 (formula search).

For task 2, we have prepared three MIR systems:

1. Soft Cosine Measure (SCM),
2. Formula2Vec, and
3. Ensemble.
Methods

Math Representations

- In our MIR systems, we used the following math representations:

- LaTeX
- Presentation MathML
- Content MathML
- Symbol Layout Tree
- M-Terms
- Operator Tree
- Prefix
- Infix
Methods

For training, we used the following two corpora:

1. ArXMLiv (1.4M articles, subsets: no_problem, warning_1, warning_2, and error), [4] and
2. Math StackExchange (2.5M posts).

For validation, we used the following two sets of relevance judgements:

1. Automatic (797K topics, 1.4M judgements, param. optimization, model selection), and
2. Human-Annotated (45 topics, 12.1K judgements, performance estimation).

In our evaluation, we used the nDCG’ [12] measure.

For retrieval, we used a machine with with 32 CPUs and 252 GiB RAM.

For training embeddings, we used an NVIDIA GTX2080 Ti GPU with 11 GiB VRAM.
Soft Cosine Measure (SCM)

- Uses TF-IDF with the Soft Cosine Measure (SCM) [13, 7, 8] document similarity.
- Uses the optimal parameters of fastText and the SCM from task 1.
- Four different fastText models were trained:
  1. Tiny (5 epochs, alternative submission)
  2. Small (10 epochs, primary submission)
  3. Medium (2 epochs on all corpora)
  4. Large (10 epochs on all corpora)

- **Accuracy**: nDCG’ 0.119 (tiny), insignificantly below the third TangentCFT+.
- **Speed**: avg. 108.86 s/topic, min. 54.81 s/topic, max. 2720.14 s/topic.
**Task 2**  
Formula2Vec

**Uses** Doc2Vec DBOW [5] and the cosine document similarity.  
**Uses** the optimal parameters of Doc2Vec from task 1.  
**Four different** Doc2Vec models were trained:  
- Tiny (5 epochs, alternative submission)  
- Small (10 epochs, primary submission)  
- Medium (2 epochs on all corpora)  
- Large (10 epochs on all corpora)  

- **Accuracy:** nDCG’ 0.108 (small), insignificantly below the third TangentCFT+.  
- **Speed:** avg. 164.5 s/topic, min. 61.61 s/topic, max. 5448.65 s/topic.
Ensemble

- Interleaves the result lists of primary submissions: SCM and Formula2Vec.
- Uses a parameter-free ensembling algorithm that only uses ranks, not scores.
- Results are ranked by median rank, then by frequency, and then interleaved.
- **Tie-breaking:** More than 50% of all results were arbitrarily interleaved.
- **Accuracy:** nDCG' 0.100, not significantly worse than the SCM. The ensemble of all non-baseline prim. submissions (0.327) not sig. worse than the second TangentCFT.
Results

- **Accuracy**: Ensemble (0.327) significantly better than SCM.

<table>
<thead>
<tr>
<th></th>
<th>SCM</th>
<th>F2Vec</th>
<th>Ens.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best²</td>
<td>0.119</td>
<td>0.108</td>
<td><strong>0.327</strong></td>
</tr>
<tr>
<td>Primary</td>
<td><strong>0.059</strong></td>
<td><strong>0.108</strong></td>
<td></td>
</tr>
<tr>
<td>Alternative</td>
<td><strong>0.119</strong></td>
<td>0.077</td>
<td><strong>0.100</strong></td>
</tr>
</tbody>
</table>

- **Speed**: Unlike in task 1, Formula2Vec slower than SCM due to dense matrix ops.

<table>
<thead>
<tr>
<th></th>
<th>SCM</th>
<th>F2Vec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td><strong>54.81</strong></td>
<td>61.61</td>
</tr>
<tr>
<td>Average</td>
<td><strong>108.86</strong></td>
<td>164.5</td>
</tr>
<tr>
<td>Maximum</td>
<td><strong>2720.14</strong></td>
<td>5448.65</td>
</tr>
</tbody>
</table>

²Includes unsubmitted out-of-competition results.
Conclusion and Future Work

- We have introduced two MIR systems:
  1. TF-IDF-based SCM, and
  2. Doc2Vec-based Formula2Vec.

- TF-IDF-based systems combine high accuracy, speed, [7] and interpretability.
- Doc2Vec-based systems provide robust performance across many tasks.
- Three is better than one: ensemble of primary submissions third in competition.
Bibliography
Bibliography


MUNI

FACULTY

OF INFORMATICS