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]

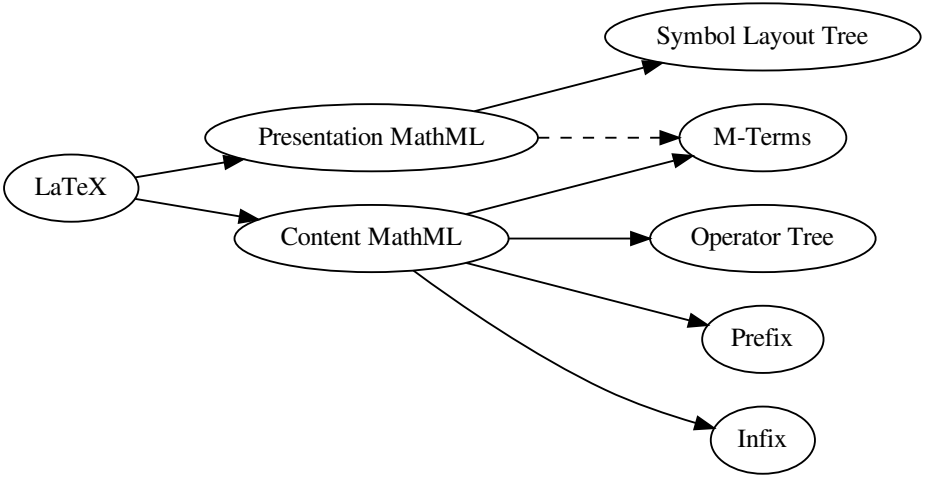
- In ARQMath 2020, we have tacked both task 1 (find answers) and 2 (formula search).
- For task 1, we have prepared five MIR systems:

  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:

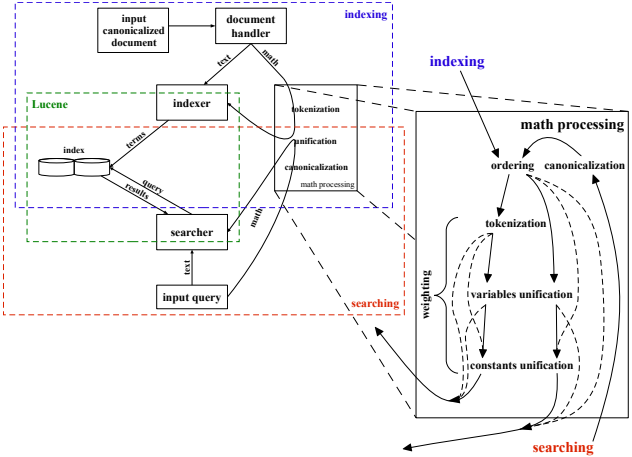


Methods

Corpora, Relevance Judgements, and Evaluation Measures

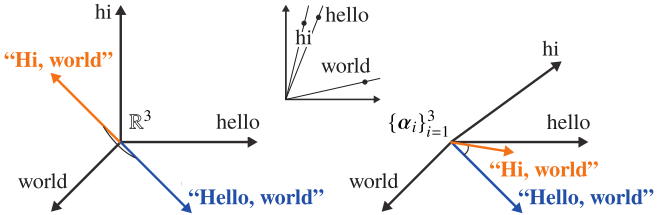
- For training, we used the following two corpora:
  1. ArXMLiv (four different subsets), [4] and
  2. Math StackExchange.
- For validation, we used the following two sets of relevance judgements:
  1. Automatic (param. opt., model sel.), and
  2. Human-Annotated (perf. est.).
- In our evaluation, we used the following two measures:
  1. Normalized Discounted Cumulative Gain Prime (nDCG'), [12] and
  2. Spearman's Correlation Coefficient (ρ).
- 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.

Math Indexer and Searcher (MIaS)



- 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)

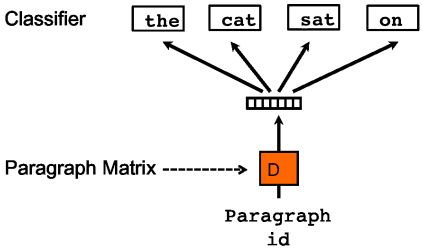


- Uses joint fastText [2] word embeddings of text & math to measure relatedness.
- Uses TF-IDF with the Prefix math representation and SCM [13, 7, 8] doc. similarity.
- Uses automatic relevance judgements to optimize parameters of fastText and SCM.
- Four different fastText models were trained:

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.

Formula2Vec

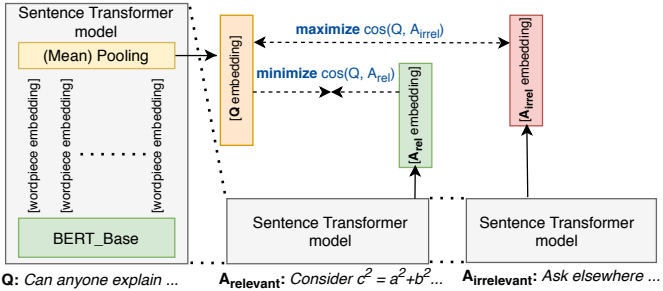


- Uses Doc2Vec DBOW [5] with the Prefix math representation and cosine doc. sim.
- Uses the optimal parameters of fastText and RedHat defaults for Doc2Vec.
- Four different Doc2Vec models were trained:

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)

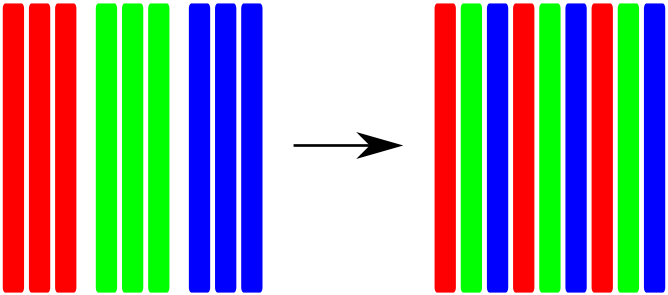
- **Accuracy:** nDCG' 0.050 (small), on par with DPRL and zbMath systems.
- **Speed:** avg. 3.23 s/topic, min. 3.14 s/topic, max. 7.87 s/topic.

CompuBERT



- 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.

	MIaS	SCM	F2Vec	CBRT	Ens.
Best[1]	0.155	*0.237*	0.101	0.009	**0.419**
Primary	*0.155*	**0.224**		0.009	
Alternative			*0.050*		**0.238**

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

	MIaS	SCM	F2Vec	CBRT
Minimum	**0.1**	30.52	*3.14*	3.2
Average	**1.24**	58.64	*3.23*	3.43
Maximum	*7.27*	502.84	7.87	**3.67**

[1] Includes 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

Bibliography I

[1] Miroslav Bartošek, Jiří Rákosník, et al. "DML-CZ: The Experience of a Medium-Sized Digital Mathematics Library". In: *Notices of the AMS* 60.8 (2013), pp. 1028–1033.

[2] Piotr Bojanowski et al. "Enriching word vectors with subword information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[3] Thierry Bouche. "Introducing EuDML-The European Digital Mathematics Library". In: *EMS Newsletter* 76 (June 2010), pp. 11–16. URL: `https://hal.archives-ouvertes.fr/hal-00559060`.

[4] Deyan Ginev. *arXMLiv:08.2019 dataset, an HTML5 conversion of arXiv.org*. SIGMathLing – Special Interest Group on Math Linguistics. 2019. URL: `https://sigmathling.kwarc.info/resources/arxmliv-dataset-082019/`.

[5] Quoc V. Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In: *CoRR* abs/1405.4053 (2014). URL: `http://arxiv.org/abs/1405.4053`.

Bibliography II

[6] Martin Líška, Petr Sojka, and Michal Růžička. "Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task". In: *Proc. of the 10th NTCIR Conference on Evaluation of Information Access Technologies*. Ed. by Noriko Kando and Kazuaki Kishida. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MATH/06-NTCIR10-MATH-LiskaM.pdf. Tokyo: NII, Tokyo, Japan, 2013, pp. 686–691. ISBN: 978-4-86049-062-1.

[7] Vít Novotný. "Implementation Notes for the Soft Cosine Measure". eng. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Torino, Italy: Association for Computing Machinery, 2018, pp. 1639–1642. ISBN: 978-1-4503-6014-2. DOI: 10.1145/3269206.3269317.

[8] Vít Novotný et al. *Text classification with word embedding regularization and soft similarity measure*. 2020. arXiv: 2003.05019 [cs.IR]. URL: https://arxiv.org/abs/2003.05019.

Bibliography III

[9] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://www.aclweb.org/anthology/D19-1410.

[10] Michal Růžička, Petr Sojka, and Martin Líška. “Math Indexer and Searcher under the Hood: Fine-Tuning Query Expansion and Unification Strategies”. eng. In: *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*. Ed. by Noriko Kando et al. Tokyo: National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan, 2016, pp. 331–337. ISBN: 978-4-86049-071-3.

Bibliography IV

[11] Michal Růžička, Petr Sojka, and Martin Líška. "Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy". In: *Proc. of the 11th NTCIR Conference on Evaluation of Information Access Technologies*. Ed. by Hideo Joho and Kazuaki Kishida. NII, Tokyo, Japan, Dec. 2014, pp. 127–134.

[12] Tetsuya Sakai and Noriko Kando. "On information retrieval metrics designed for evaluation with incomplete relevance assessments". In: *Information Retrieval* 11.5 (2008), pp. 447–470.

[13] Grigori Sidorov et al. "Soft similarity and soft cosine measure: Similarity of features in vector space model". In: *Computación y Sistemas* 18.3 (2014), pp. 491–504.

[14] Krzysztof Wojciechowski et al. *The EuDML Search and Browsing Service – Final*. Deliverable D5.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, revision 1.2 `https://project.eudml.eu/sites/default/files/D5_3_v1.2.pdf`. Feb. 2013.

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]

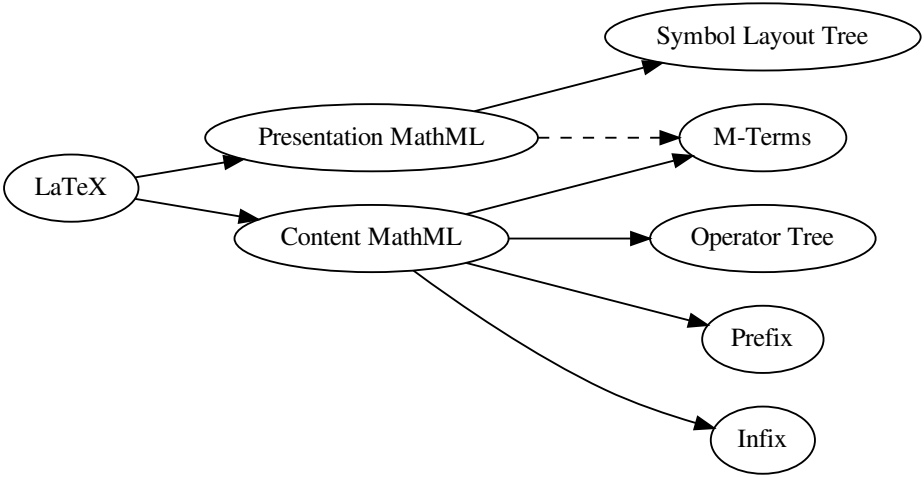
- In ARQMath 2020, we have tacked both task 1 (find answers) and 2 (formula search).
- For task 2, we have prepared three MIR systems:

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

Methods

Math Representations

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



Methods

- For training, we used the following two corpora:

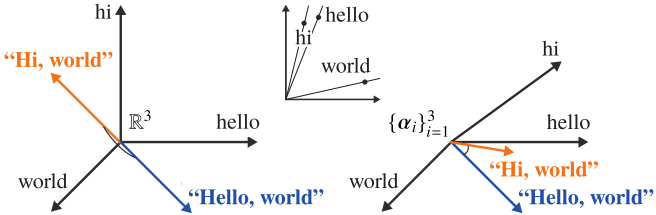
  1. ArXMLiv (1.4M articles, subsets: no_problem, warning_1, warning_2, and error), [4] and
  2. Math StackExchange (2.5M posts).

- 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)

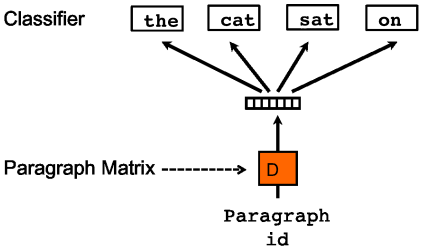


- Uses joint fastText [2] word embeddings of text & math to measure token similarity.
- Uses TF-IDF with the Soft Cosine Measure (SCM) [13, 7, 8] document similarity.
- Uses the optimal parameters of fastText and the SCM from task 1.
- Four different fastText models were trained:

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.

Formula2Vec

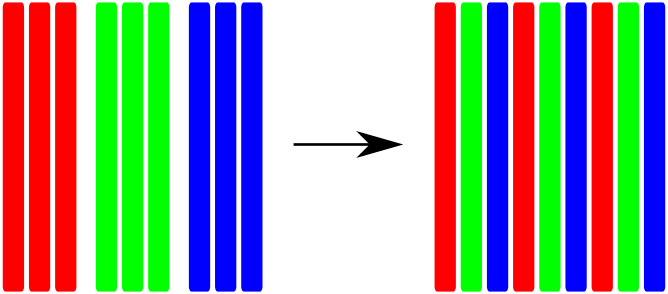


- Uses Doc2Vec DBOW [5] and the cosine document similarity.
- Uses the optimal parameters of Doc2Vec from task 1.
- Four different Doc2Vec models were trained:

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' 0.108 (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.

	SCM	F2Vec	Ens.
Best[2]	*0.119*	0.108	**0.327**
Primary	*0.059*	**0.108**	
Alternative	**0.119**	0.077	*0.100*

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

	SCM	F2Vec
Minimum	**54.81**	*61.61*
Average	**108.86**	*164.5*
Maximum	**2720.14**	*5448.65*

[2] Includes 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 I

[1] Miroslav Bartošek, Jiří Rákosník, et al. "DML-CZ: The Experience of a Medium-Sized Digital Mathematics Library". In: *Notices of the AMS* 60.8 (2013), pp. 1028–1033.

[2] Piotr Bojanowski et al. "Enriching word vectors with subword information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

[3] Thierry Bouche. "Introducing EuDML-The European Digital Mathematics Library". In: *EMS Newsletter* 76 (June 2010), pp. 11–16. URL: `https://hal.archives-ouvertes.fr/hal-00559060`.

[4] Deyan Ginev. *arXMLiv:08.2019 dataset, an HTML5 conversion of arXiv.org*. SIGMathLing – Special Interest Group on Math Linguistics. 2019. URL: `https://sigmathling.kwarc.info/resources/arxmliv-dataset-082019/`.

[5] Quoc V. Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In: *CoRR* abs/1405.4053 (2014). URL: `http://arxiv.org/abs/1405.4053`.

Bibliography II

[6] Martin Líška, Petr Sojka, and Michal Růžička. "Similarity Search for Mathematics: Masaryk University team at the NTCIR-10 Math Task". In: *Proc. of the 10th NTCIR Conference on Evaluation of Information Access Technologies*. Ed. by Noriko Kando and Kazuaki Kishida. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/pdf/NTCIR/MATH/06-NTCIR10-MATH-LiskaM.pdf. Tokyo: NII, Tokyo, Japan, 2013, pp. 686–691. ISBN: 978-4-86049-062-1.

[7] Vít Novotný. "Implementation Notes for the Soft Cosine Measure". eng. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*. Torino, Italy: Association for Computing Machinery, 2018, pp. 1639–1642. ISBN: 978-1-4503-6014-2. DOI: 10.1145/3269206.3269317.

[8] Vít Novotný et al. *Text classification with word embedding regularization and soft similarity measure*. 2020. arXiv: 2003.05019 [cs.IR]. URL: https://arxiv.org/abs/2003.05019.

Bibliography III

[9] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410. URL: https://www.aclweb.org/anthology/D19-1410.

[10] Michal Růžička, Petr Sojka, and Martin Líška. “Math Indexer and Searcher under the Hood: Fine-Tuning Query Expansion and Unification Strategies”. eng. In: *Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies*. Ed. by Noriko Kando et al. Tokyo: National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430 Japan, 2016, pp. 331–337. ISBN: 978-4-86049-071-3.

Bibliography IV

[11] Michal Růžička, Petr Sojka, and Martin Líška. "Math Indexer and Searcher under the Hood: History and Development of a Winning Strategy". In: *Proc. of the 11th NTCIR Conference on Evaluation of Information Access Technologies*. Ed. by Hideo Joho and Kazuaki Kishida. NII, Tokyo, Japan, Dec. 2014, pp. 127–134.

[12] Tetsuya Sakai and Noriko Kando. "On information retrieval metrics designed for evaluation with incomplete relevance assessments". In: *Information Retrieval* 11.5 (2008), pp. 447–470.

[13] Grigori Sidorov et al. "Soft similarity and soft cosine measure: Similarity of features in vector space model". In: *Computación y Sistemas* 18.3 (2014), pp. 491–504.

[14] Krzysztof Wojciechowski et al. *The EuDML Search and Browsing Service – Final*. Deliverable D5.3 of EU CIP-ICT-PSP project 250503 EuDML: The European Digital Mathematics Library, revision 1.2 `https://project.eudml.eu/sites/default/files/D5_3_v1.2.pdf`. Feb. 2013.

MUNI
FACULTY
OF INFORMATICS