Math-aware Similarity of Papers in Digital Mathematics Libraries

Petr Sojka

Masaryk University, Faculty of Informatics, Czech Republic
sojka@fi.muni.cz

The talk is based on the joint work with Radim Řehůřek and Michal Růžička (Masaryk University Brno).

Session: 19. Information and Communication in Mathematics

The exploratory, semantic similarity searching is becoming widespread in digital libraries, and math ones are no exception. For working mathematicians and their use of digital mathematical libraries (DML) as the Czech Digital Mathematics Library DML-CZ [1] or European Digital Mathematics Library (EuDML) [2] we have designed and implemented math-aware similarity computation framework based on leading edge topic modelling techniques implemented by Gensim software package [3].

Studies on the classification of math papers done for DML-CZ [4] have been tested and deployed in EuDML, where for given paper ten most semantically similar papers are computed and shown. In the latest experiments we are evaluating several possible representations of mathematical formulae to get the *semantically* similar papers. Quality of similarity is measured by comparation to the similarity matrix induced from the Mathematical Subject Classifications every paper is marked up by.

In the talk we will report **a)** about the evaluation of the similarities computed by several different methods, **b)** on the experience from 20 months of deployment in EuDML and more than 5 years in DML-CZ, **c)** about the importance of representing formulae even for paper similarity computations, **d)** on setting up Gensim for the math-aware use in DML projects.

References

[1] P. Sojka, J. Rákosník, *From Pixels and Minds to the Mathematical Knowledge in Digital Library*, In P. Sojka (ed.): Proceedings of DML 2008: Towards a Digital Mathematics Library. Brno: Masaryk University, 2008. pp. 17–27, https://is.muni.cz/publication/762453?lang=en.

[2] J. Borbinha, T. Bouche, A. Nowiński, P. Sojka, *Project EuDML–A First Year Demonstration*, In J.H. Davenport et al. (eds.): Proceedings of CICM 2011, Springer, LNAI vol. 6824, 2011. pp. 281–284, doi:10.1007/978-3-642-22673-1_21.

[3] R. Řehůřek, P. Sojka, *Software Framework for Topic Modelling with Large Corpora*. In Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks. Valletta, Malta: University of Malta, 2010. pp. 46–50, http://is.muni.cz/publication/884893/en.

[4] R. Řehůřek, P. Sojka, *Automated Classification and Categorization of Mathematical Knowledge*, In S. Autexier et al. (eds.): Proceedings of CICM 2008, Springer, LNAI vol. 5144, 2008. pp. 543–557, 10.1007/978-3-540-85110-3_44