Automatic classification and clustering of mathematical publications

Simon Barthel

L3S, Germany
s.barthel@tu-bs.de

Session: 19. Information and Communication in Mathematics

Since the 1940s the Mathematics Subject Classification (MSC) have been used to categorize the whole area of mathematical publications. The MSC is a manually created taxonomy which is in a contant process of manual refactoring. The current version of the MSC contains 63 on the top-level, 528 classes on the second level and 5606 leaf nodes. Currently, every document that is published on the ZentralblattMATH as well as on the Mathematical Reviews is manually categorized with respect to the MSC with a huge amount of effort.

It seems obvious that this process can be easily automated by using modern mashine learning approaches, but first experiments showed that state-of-the-art mashine learning algorithms that are based on traditional bag-of-word vector-space-models are not capable to succeed in this task. This fact raises the assumption that texts from different mathematical disciplines are not necessarily distinguishable based on text features.

In this talk we will therefore analyze how mashine learning algorithms would cluster mathematical documents and how this automatically generated clusters can be characterized. We will then compare this automatically generated clusters to the manually created MSC and discuss the differences.