Crowdsourcing the reconstruction of the Cairo Genizah

Lior Wolf, Tel-Aviv University

Many significant historical corpora contain leaves that are mixed up and no longer bound in their original state as multi-page documents. The reconstruction of old manuscripts from a mix of disjoint leaves can therefore be of a paramount importance to historians and literary scholars. Previously, it was shown that visual similarity provides meaningful pair-wise similarities between handwritten leaves. Here, we go a step further and suggest a semiautomatic clustering tool that helps reconstruct the original documents. The proposed solution is based on a graphical model that makes inferences based on catalog information provided for each leaf as well as on the pairwise similarities of handwriting. Several novel active clustering techniques

are explored, and the solution is applied to a significant part of the Cairo Genizah, where the problem of joining leaves remains unsolved even after a century of extensive study by hundreds of human scholars.