System and Method for Document Analysis, Processing and Information Extraction

**OCR Number:** OCR 4683

**Description:**

The present invention is directed to a method and computer system for representing a dataset comprising N documents by computing a diffusion geometry of the dataset comprising at least a plurality of diffusion coordinates. The present method and system stores a number of diffusion coordinates, wherein the number is linear in proportion to N.

**Licensing Contact:** Richard Andersson  
richard.andersson@yale.edu