



# A Method to extend Existing Document Clustering Procedures by including Relational Information

Tijn Witsenburg<sup>1</sup>

Hendrik Blockeel<sup>1,2</sup>

Universiteit Leiden

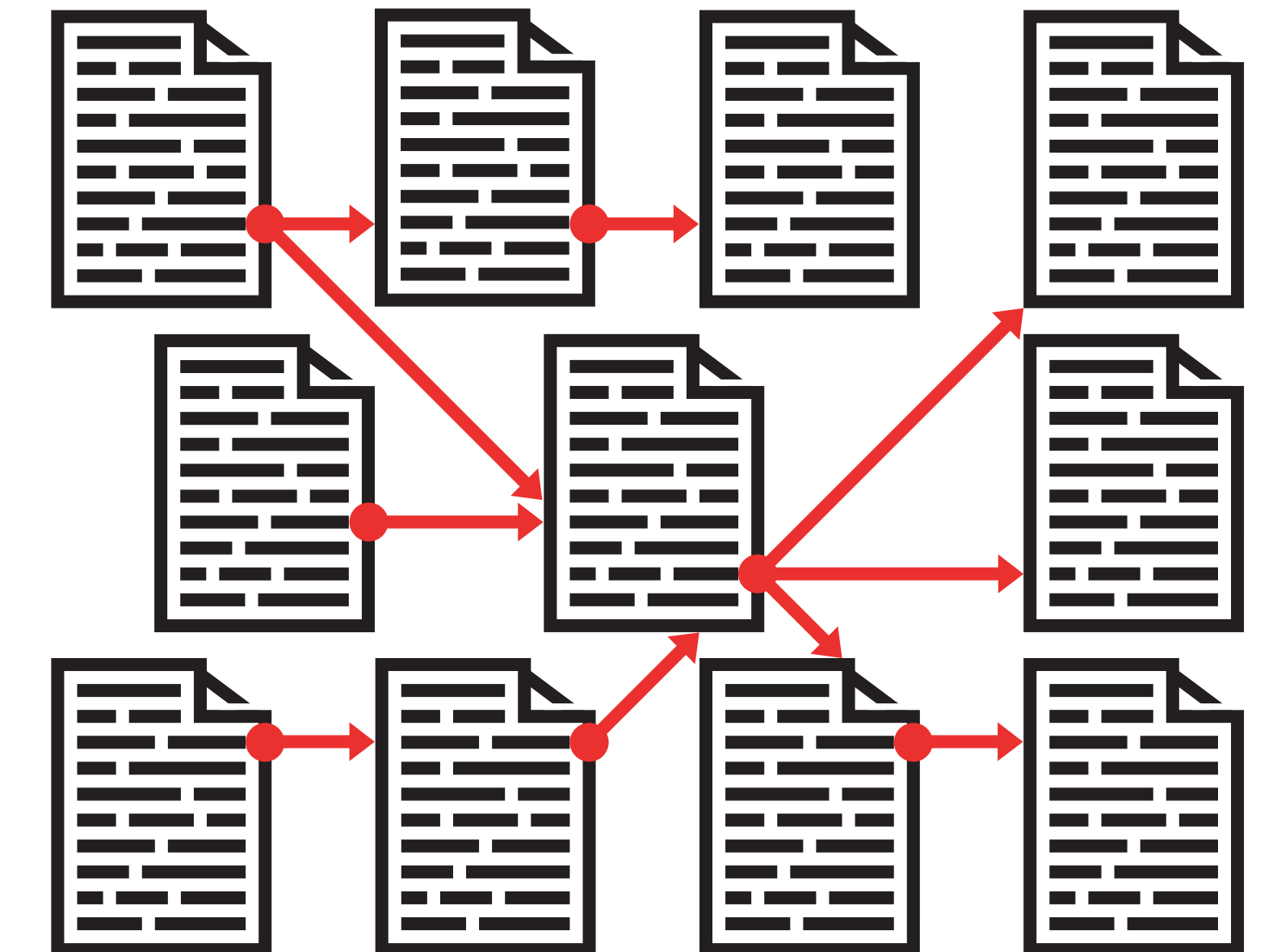
<sup>1</sup> Leiden Institute of Advanced Computer Science, Universiteit Leiden

<sup>2</sup> Department of Computer Science, Katholieke Universiteit Leuven

## Problem Setting

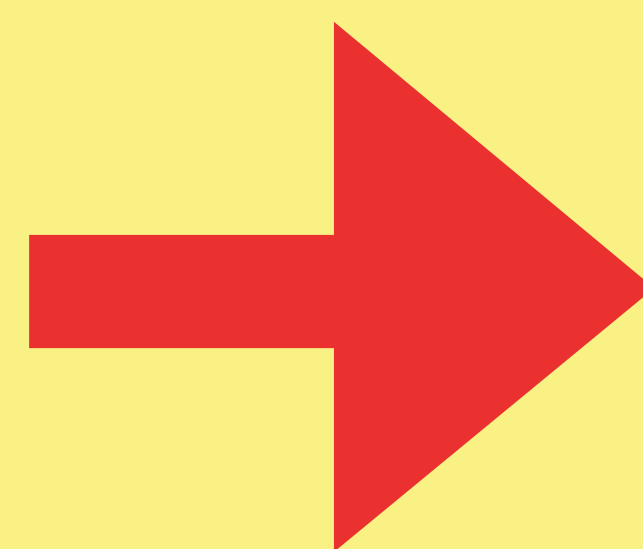
Consider data that consist of both content and relational information. Can existing clustering procedures that use content be improved by inserting relational information?

*for instance scientific papers with citations or common authors* ▶



## Common approach in Clustering

1. Translation of data to numeric values.
2. Use values to calculate distances / similarities.
3. Use these to divide data items into clusters.



## Basic Idea of our Method

Before dividing the data items into clusters, the distance / similarity matrix is altered to include the relational information.

## Altering the Distance/Similarity Matrix

First a symmetric adjacency matrix  $A$  is created. Then  $A$  is combined with the distance / similarity matrix  $M$  in a way that is derived from matrix multiplication.

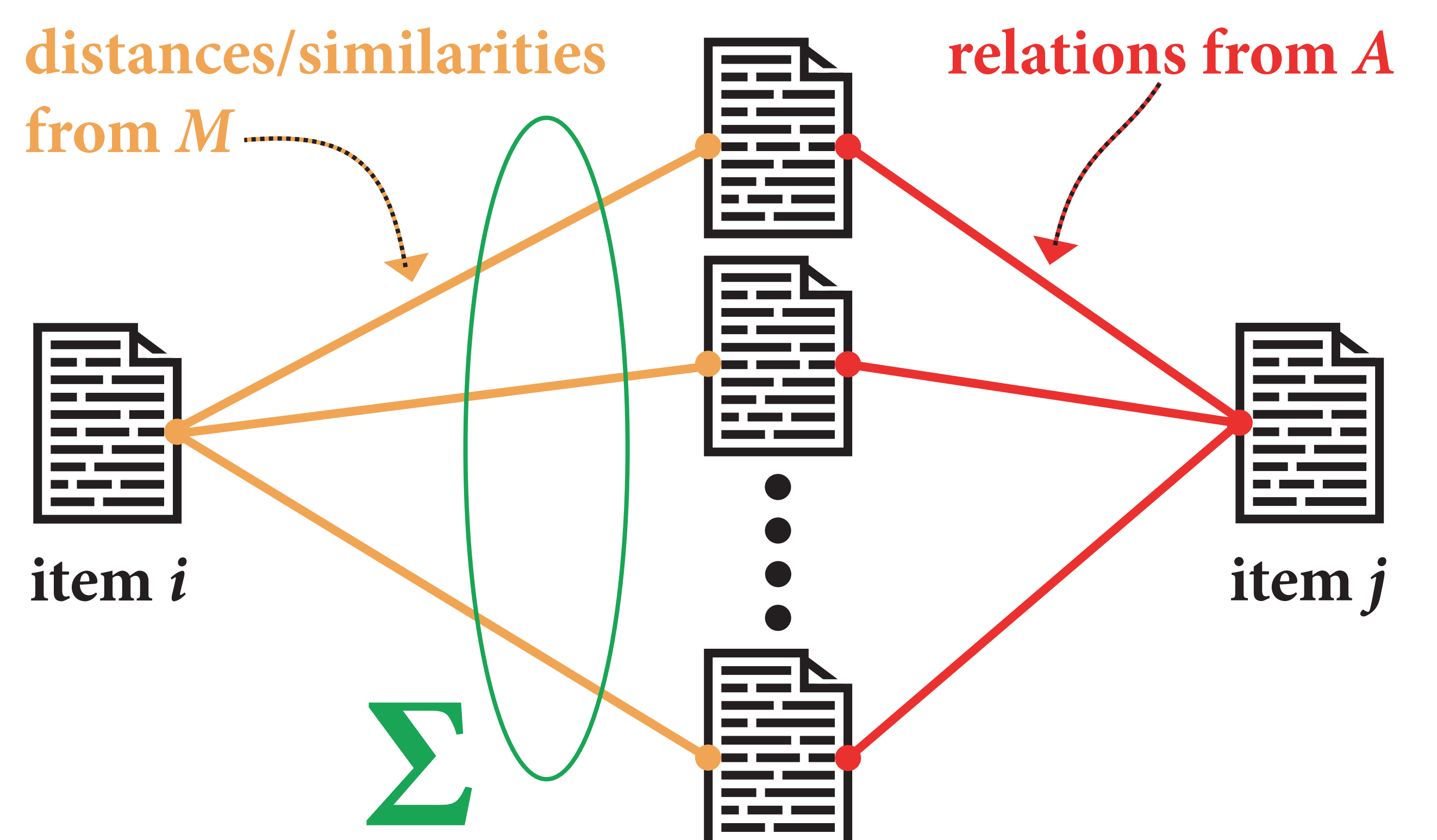
A new matrix  $M'$  is created by using the equation:

$$M' = M \times A + A \times M$$

This will give the sum as described to the right. When we do not wish to use the sum but the average value, and have  $N$  nodes, any element  $m'_{ij}$  of  $M'$  can be calculated using:

$$m'_{ij} = \frac{1}{2} \cdot \frac{\sum_{x=0}^N m_{ix} \cdot a_{xj}}{\sum_{x=0}^N a_{xj}} + \frac{\sum_{x=0}^N a_{ix} \cdot m_{xj}}{\sum_{x=0}^N a_{ix}}$$

## Meaning of element $m'_{ij}$ in $M \times A$



Every element  $m'_{ij}$  from  $M' = M \times A$  can be seen as the sum of the distances/similarities between  $i$  and all nodes that have a relation with  $j$ .

## Preliminary Results

This method has been tested on sub sets of a single database using a simple, greedy clustering procedure. First results look promising, but further research needs to be done.

## Questions:

- Could this method be used on **other databases**?
- Will this method work for **other clustering procedures**?
- How can **other information** be included as well?
- Can this principle be used on **other research fields**?
- Can anyone *help me finding other good databases*?

