Revealing Relations between Open and Closed Answers in Questionnaires through Text Clustering Evaluation

Magnus Rosell and Sumithra Velupillai

2008-05-28

Content

- Introduction and Motivation
- Text Set and Exploration Tool
- Method
- Example
- Evaluation
- Conclusion

Introduction and Motivation

- Questionnaires – an important source for research
- Hidden information in open free-text answers
- Time-consuming to analyze manually
- Text clustering could aid

Text Set and Exploration Tool

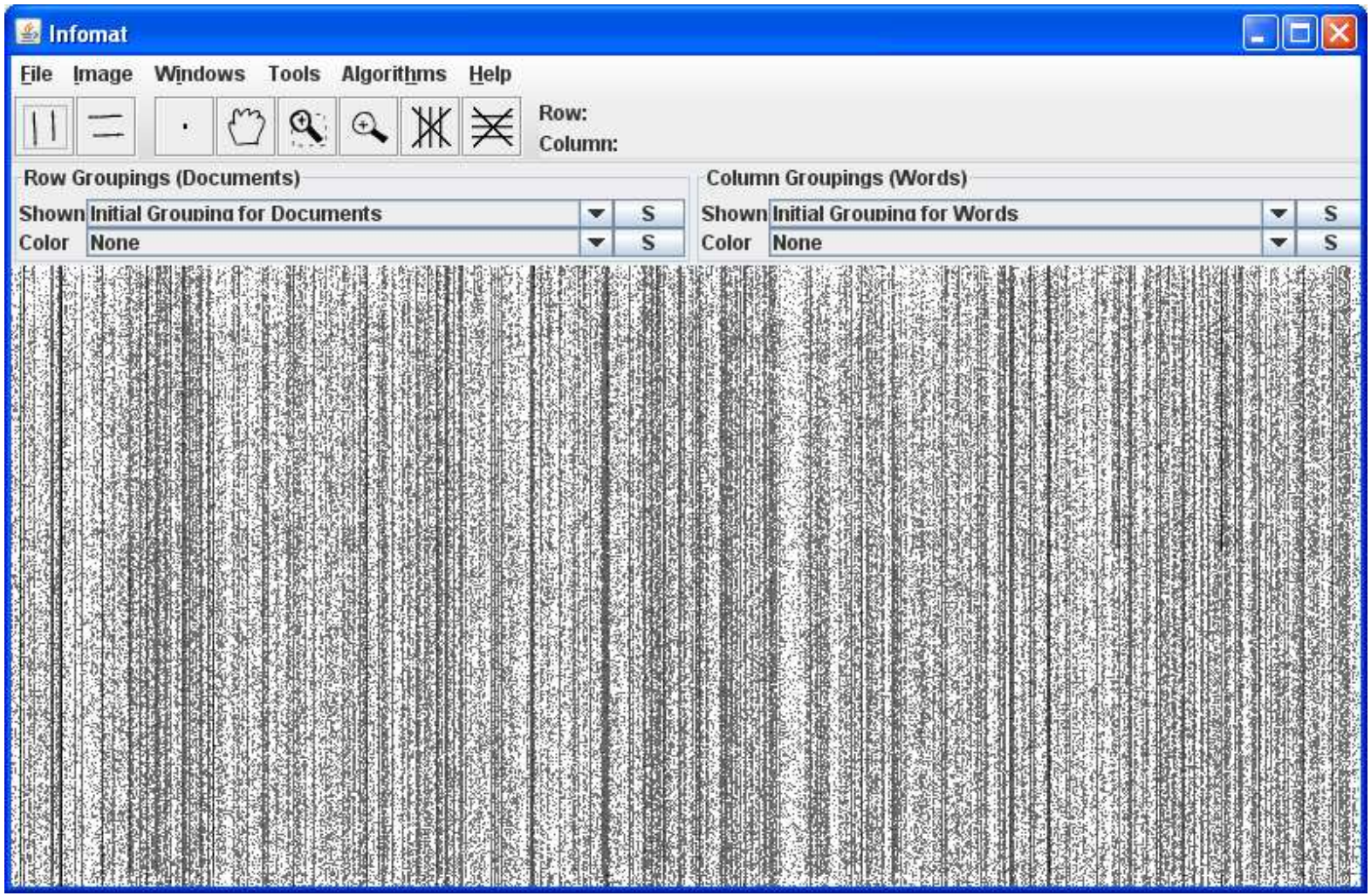
- The Swedish Twin Registry
- A questionnaire
  - An open answer: occupation (41 549)
    - Vector space model
  - A closed answer: smokers (29%)
- Infomat – A vector space exploration tool
  - `http://www.csc.kth.se/tcs/humanlang/tools.html`

Method

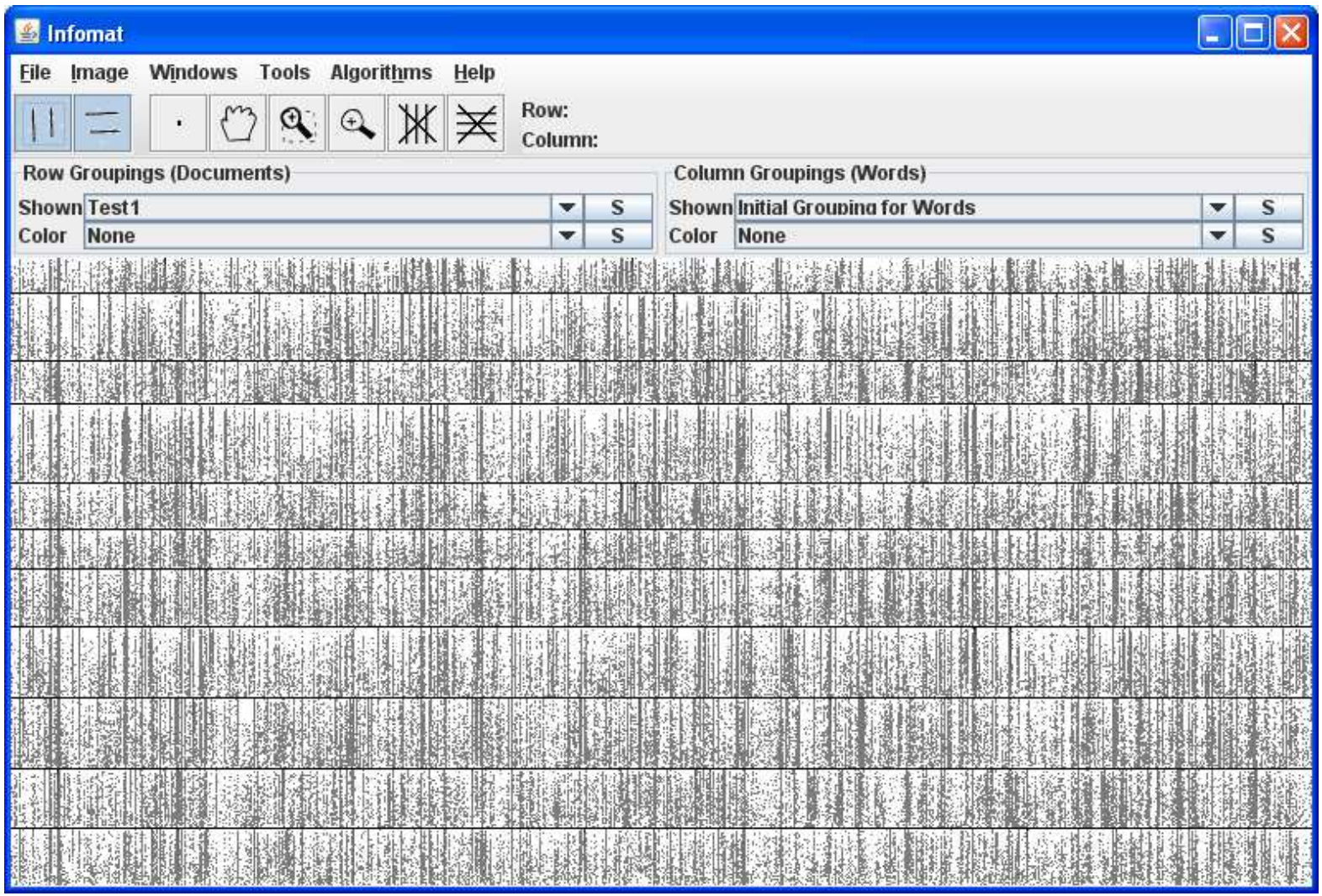
1. Cluster the text set
2. Identify interesting clusters
3. Explore cluster contents
4. Formulate potential hypotheses

- Iterate
- Interactive exploration
- Pursue hypotheses further

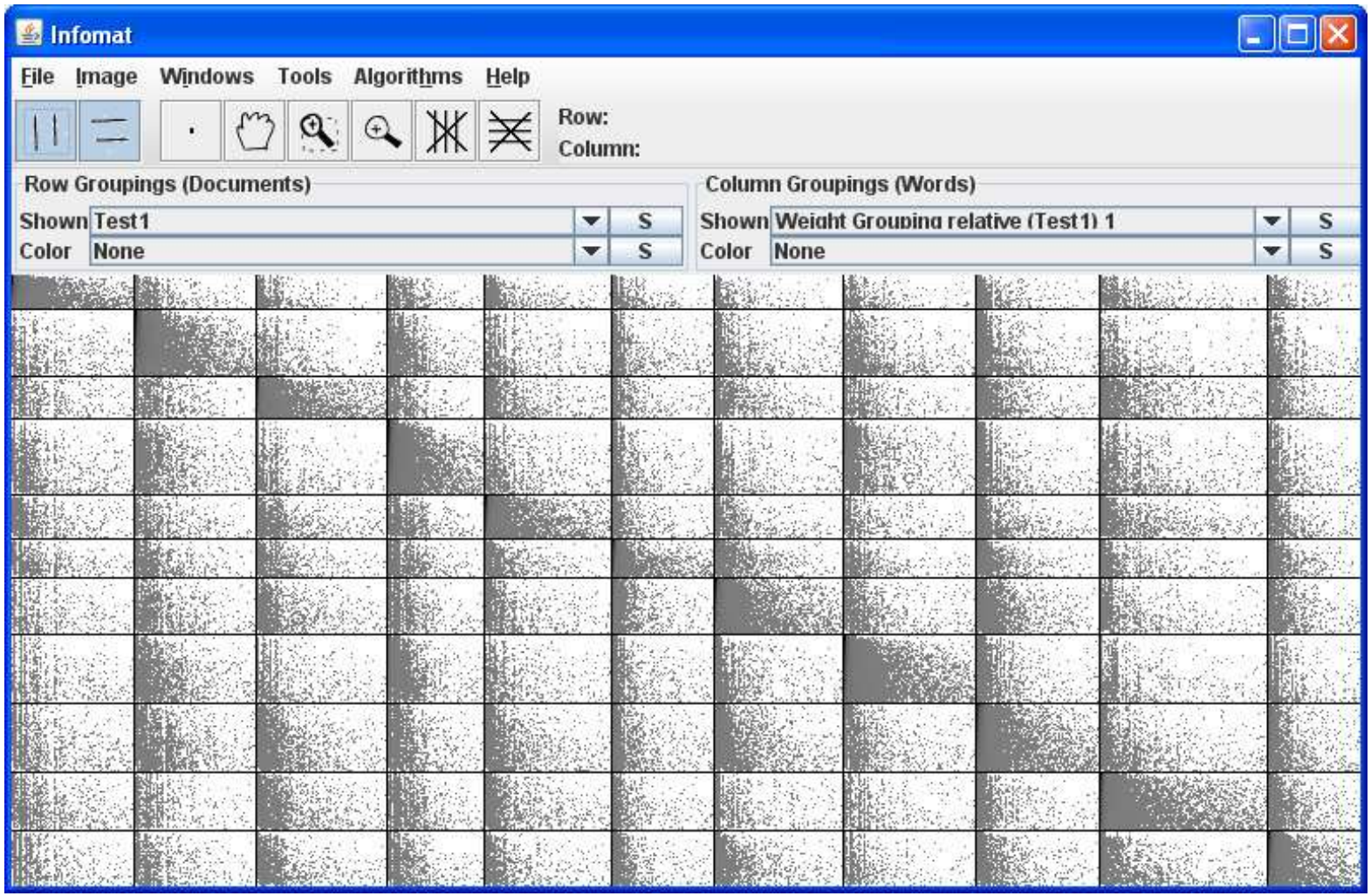
Example – Text Set



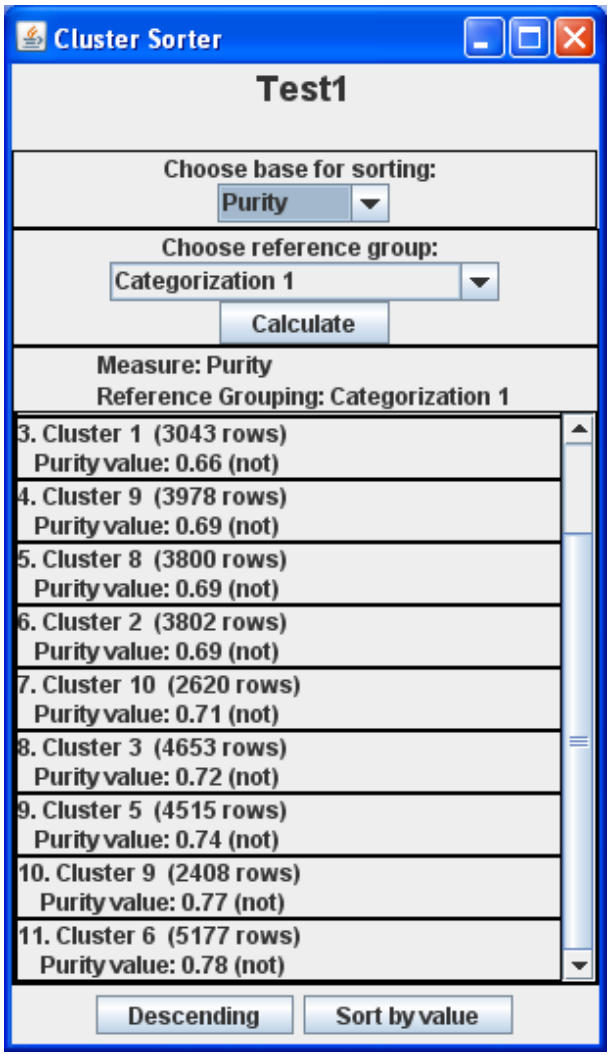
Example – Clustering



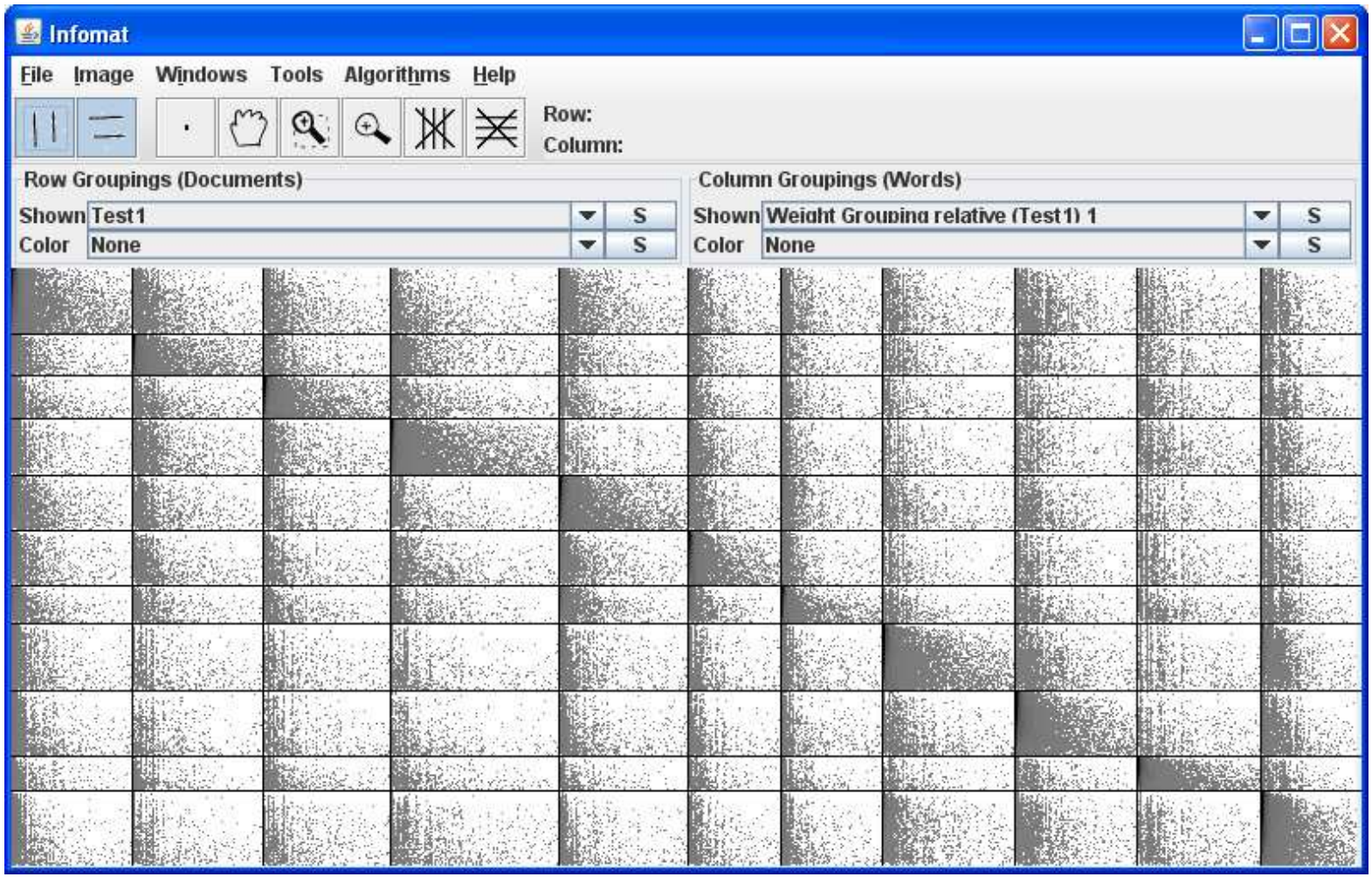
Example – Relative Clustering



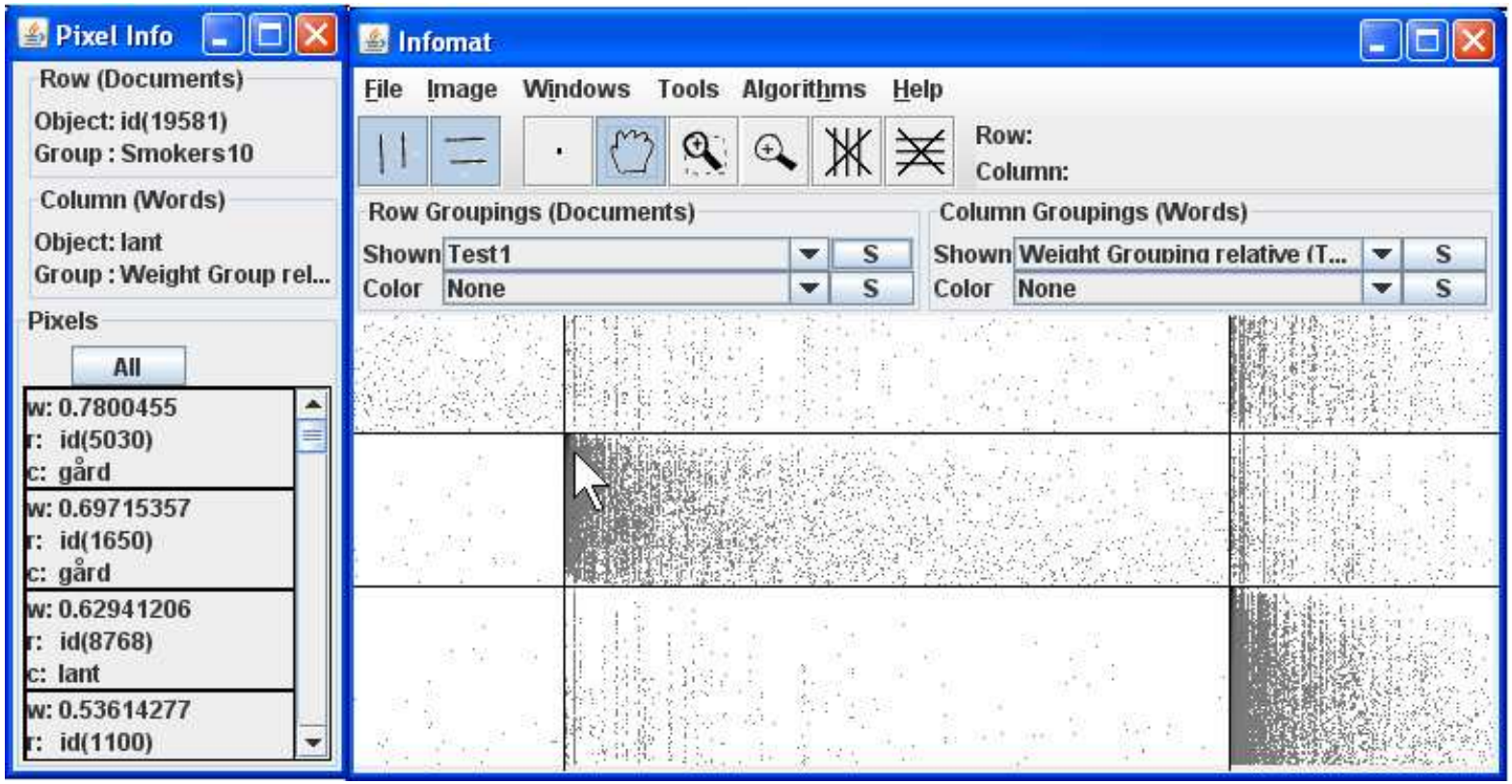
Example – Evaluated Clusters



Example – Sorted Clustering



Example – Zoom on an Interesting Cluster



Evaluation

- Farmers smoke less than the average
  - A few hours of exploration
  - No prior knowledge on smoking habits in occupation groups
- Comparable surveys
- Hypotheses can be generated

Conclusion

- No need to avoid free-text answers
  - valuable
  - analyze with our method
- Previously unknown relations were revealed
- Interaction!
- Pursue hypothesis further
- Future work
  - Questionnaires from other domains
  - Similar text sets, e.g. electronic medical records