Cluster Analysis Applied to Europana

Esra Atescelik

Supervisors

Victor de Boer, Antoine Isaac

Institutions

VU University Amsterdam, Europeana

Contents

- Context & Problem Statement
- Research Questions
- Approach
- Experiments
  - Experimental Setup
  - Evaluation results
- Conclusion & Discussion
- Future Work

Context & Problem Statement

- Europeana gives people free and open access to millions of digital objects (e.g. books, paintings, films, sounds etc.) throughout Europe

- Europeana does not maintain the collection-level metadata yet

Example of metadata for an object on Europeana

© Free access - no re-use

View item at
Sinebrychoffin taidemuseo

Before and after I

Creator: Hogarth, William (taiteilija)

Date of creation: 1736

Type: Physical Object ; grafiikka

Format: kuvan mitat 37 x 30,2 cm ; lehden mitat 38,5 x 30,8 cm ; etsaus

Identifier: RAMSAY 506:1

Source: Sinebrychoffin taidemuseo

Data provider: Sinebrychoffin taidemuseo

Provider: National Library of Finland

Providing country: Finland

Auto-generated tags

Example of collection from The European Library (TEL)

Vienna

Contributor
Austrian National Library

Collection type
Collection

Search for items in the collection... GO

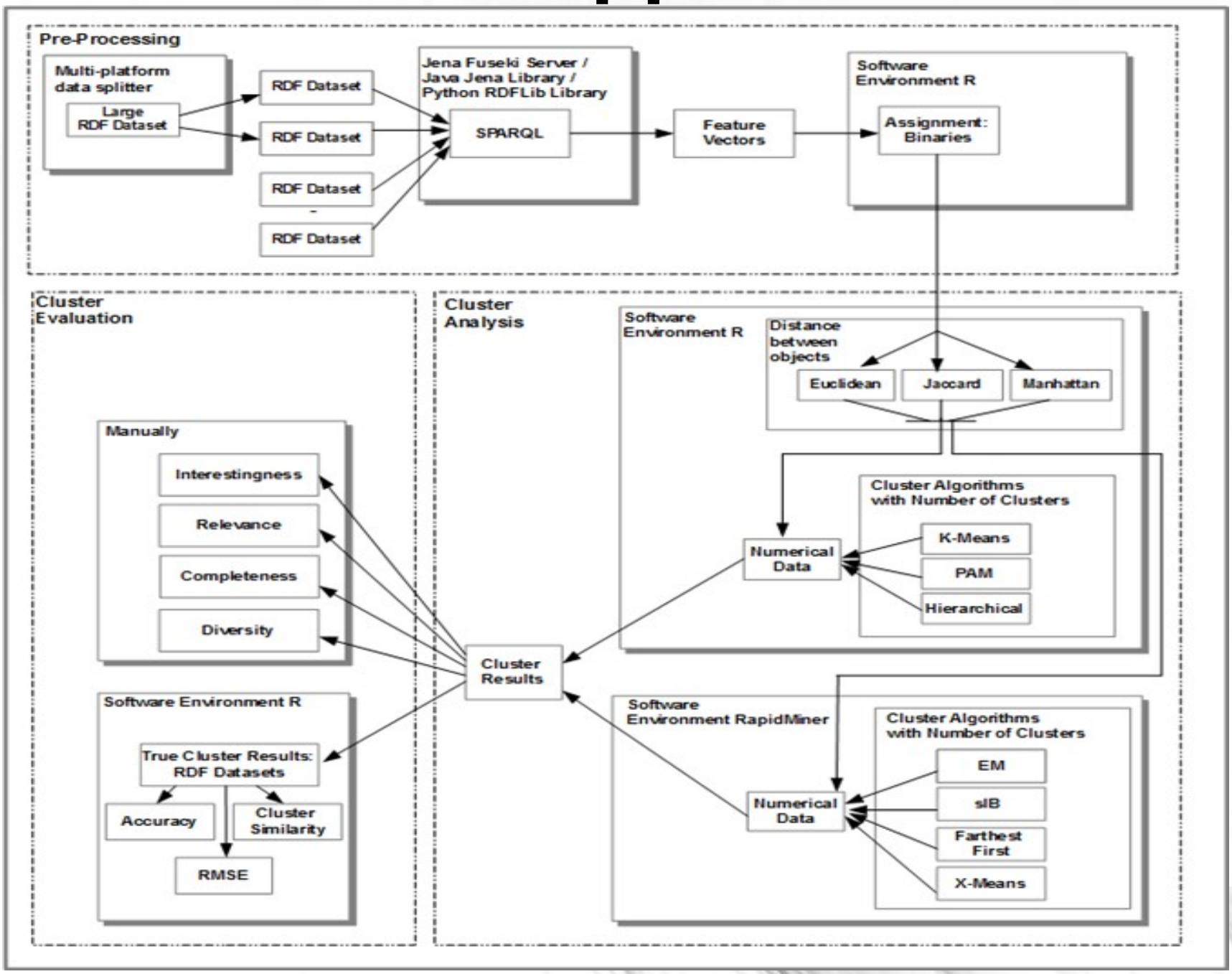
Description

This collection of old Viennese photographs covers several themes: people, Viennese history, everyday life, Viennese surroundings, places and streets of all 23 districts within Vienna. The collection has been assembled by The Picture Archive of the Austrian National Library (ANL), Austria's largest picture documentation centre and the centre for the ordering of digital images of all holdings and collections. Users can access the collection through the Austria Picture Archive, an image platform of the library. Within this Catalogue, users can also search by districts, streets, house numbers, as well as by the names of main buildings.

Research Questions

- How can we cluster the Europeana objects?
- What is the clustering method with the best performance for clustering Europeana metadata?
- What are the effects of different parameters on the cluster analysis results?

Our Approach



Pre-processing

- Data: Europeana RDF Datasets
  - Big RDF dataset →

    Multi-platform data splitter: *HJSplit*
- Extraction metadata of RDF datasets
  - *SPARQL*
- Process SPARQL on the platforms
  - Jena Fuseki Server
  - *Java Jena Library*
  - Python RDFLib Library

SPARQL - Pre-processing (II)

Fuseki Query

SPARQL Query

```
PREFIX edm: <http://www.europeana.eu/schemas/edm/>
       PREFIX ore: <http://www.openarchives.org/ore/terms/>
       PREFIX dc: <http://purl.org/dc/elements/1.1/>
       REFIX dct: <http://purl.org/dc/terms/>
       SELECT DISTINCT ?proxyProvider (GROUP_CONCAT(DISTINCT ?title) AS ?titles)
        (GROUP_CONCAT(DISTINCT ?type) AS ?types)
       (GROUP_CONCAT(DISTINCT ?format) AS ?formats)
        (GROUP_CONCAT(DISTINCT ?date) AS ?dates)
       (GROUP_CONCAT(DISTINCT ?publisher) AS ?publishers)
        (GROUP_CONCAT(DISTINCT ?relation) AS ?relations)
       (GROUP_CONCAT(DISTINCT ?subject) AS ?subjects)
        (GROUP_CONCAT(DISTINCT ?identifier) AS ?identifiers)
       (GROUP_CONCAT(DISTINCT ?source) AS ?sources)
        (GROUP_CONCAT(DISTINCT ?edmtype) AS ?edmtypes)
       (GROUP_CONCAT(DISTINCT ?creator) AS ?creators)
        (GROUP_CONCAT(DISTINCT ?contributor) AS ?contributors)
       (GROUP_CONCAT(DISTINCT ?right) AS ?rights)
        (GROUP_CONCAT(DISTINCT ?provider) AS ?providers)
       (GROUP_CONCAT(DISTINCT ?dataProvider) AS ?dataProviders)
        (GROUP_CONCAT(DISTINCT ?country) AS ?countrys)
       (GROUP_CONCAT(DISTINCT ?language) AS ?languages)
                WHERE {
                        ?proxyProvider ore:proxyIn ?aggregationProvider
                        OPTIONAL {?proxyProvider dc:title ?title }
                        OPTIONAL {?proxyProvider dc:type ?type }
                        OPTIONAL {?proxyProvider dc:extent ?format }
                        OPTIONAL {?proxyProvider dc:date ?date }
                        OPTIONAL {?proxyProvider dc:publisher ?publisher }
                        OPTIONAL {?proxyProvider dc:relation ?relation }
                        OPTIONAL {?proxyProvider dc:subject ?subject }
                        OPTIONAL {?proxyProvider dc:identifier ?identifier }
                        OPTIONAL {?proxyProvider dc:source ?source }
                        OPTIONAL {?proxyProvider edm:type ?edmtype }
                        OPTIONAL {?proxyProvider dc:creator ?creator }
                        OPTIONAL {?proxyProvider dc:contributor ?contributor }
                        OPTIONAL {?aggregationProvider dc:rights ?right }
                        OPTIONAL {?aggregationProvider edm:provider ?provider }
                        OPTIONAL {?aggregationProvider edm:dataProvider ?dataProvider
                        ?aggregationEuropeana ore:aggregates ?aggregationProvider
                                  OPTIONAL {?aggregationEuropeana edm:country ?country }
                                  OPTIONAL {?aggregationEuropeana edm:language ?language }
   } GROUP BY ?proxyProvider
```

Output: Text

If XML output, add XSLT style sheet (blank for none):

Force the accept header to text/plain regardless.

Get Results

Pre-processing (III)

- Output of metadata extraction process
  - Feature vectors
- Assign binaries to the categorical attribute values using *R*
  - Complete disjunctive table of a factor table
    - Not used in assigning of binaries:

      Non-categorical attributes, attributes with one or no value, missing attribute values

Cluster Analysis

- Binaries
  - Distance Measures using *R*
    - Euclidean
    - Jaccard
    - Manhattan
- Numerical
  - Clustering Algorithms using *RapidMiner*, *R*
    - K-Means, PAM, Hierarchical, EM, sIB
      Farthest First, X-Means
- Cluster Results
  - Table of integers e.g. Cluster labels 1, 2 etc.

Cluster Evaluation

- True Cluster Vectors
  - Number of clusters (*k*) is equal to the number of datasets we use in the clustering
    - Europeana RDF datasets as “gold standard”
    - The European Library (TEL)
  - Evaluation Metrics
    - Accuracy
    - Cluster Similarity
    - RMSE
    - Running Time

Cluster Evaluation (II)

- Manually
  - With the help of two domain experts
  - Best clustering method
  - 4 subjective factors
  - Evaluation Metrics
    - Interestingness
    - Relevance
    - Completeness
    - Diversity

Experiments

- 3 experiments with different goals
  - Experiment 1
    - Show how can we can perform the cluster analysis on Europeana datasets
  - Experiment 2
    - Find “best” parametric setting and “best” clustering algorithm
  - Experiment 3
    - Examine if the clustering algorithm is good enough for Europeana

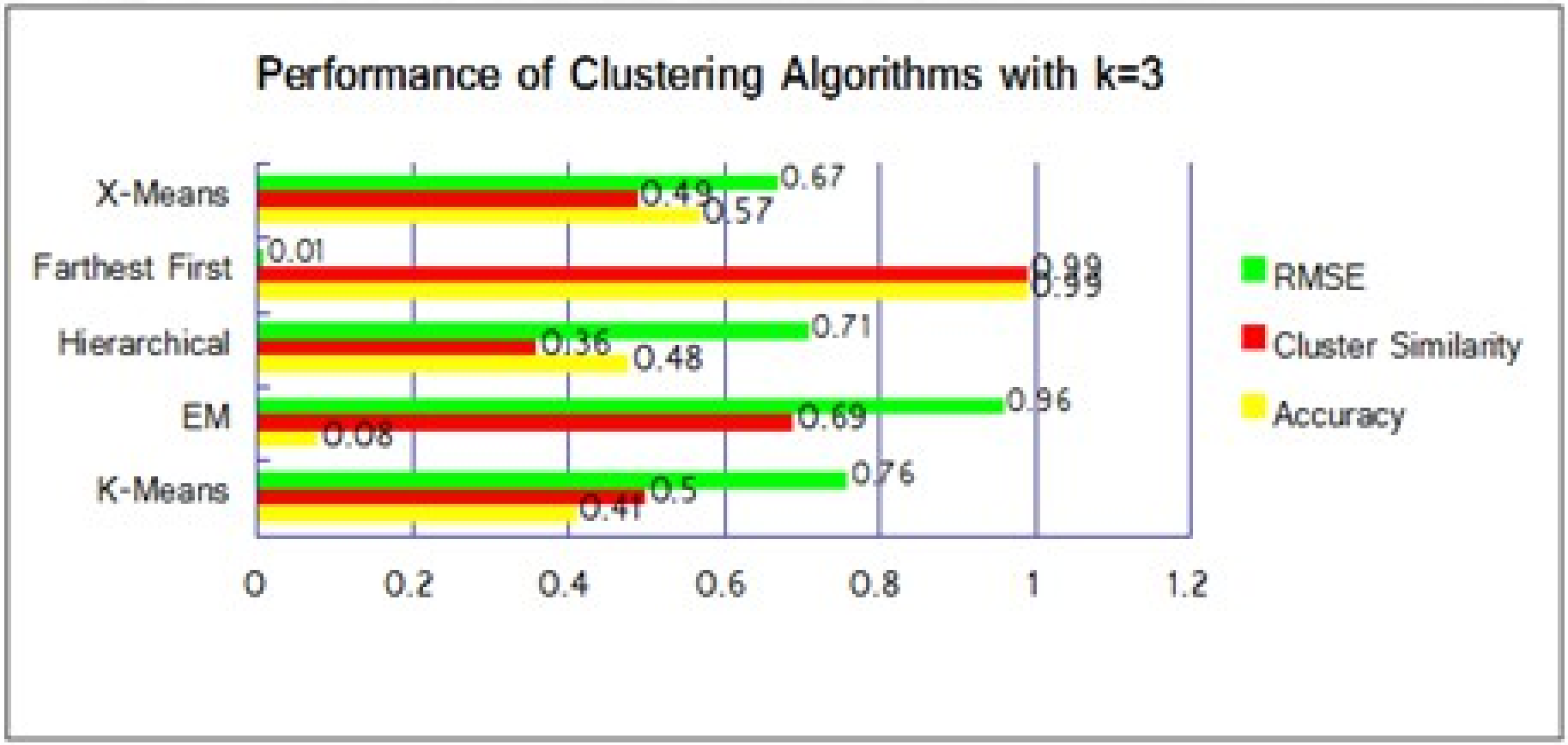
Experiments (II)

- Cluster Evaluation
  - Experiment 1 & 2
    - True Cluster Vectors
  - Experiment 3
    - Manually

Experiment 1

- 3 datasets (randomly) with the provider 'The European Library' so $k = 3$
- 9413 Europeana objects
- Clustering algorithms
  - K-Means, EM, Farthest First
    - No distance
  - Hierarchical, X-Means
    - Euclidean

Cluster Evaluation Results



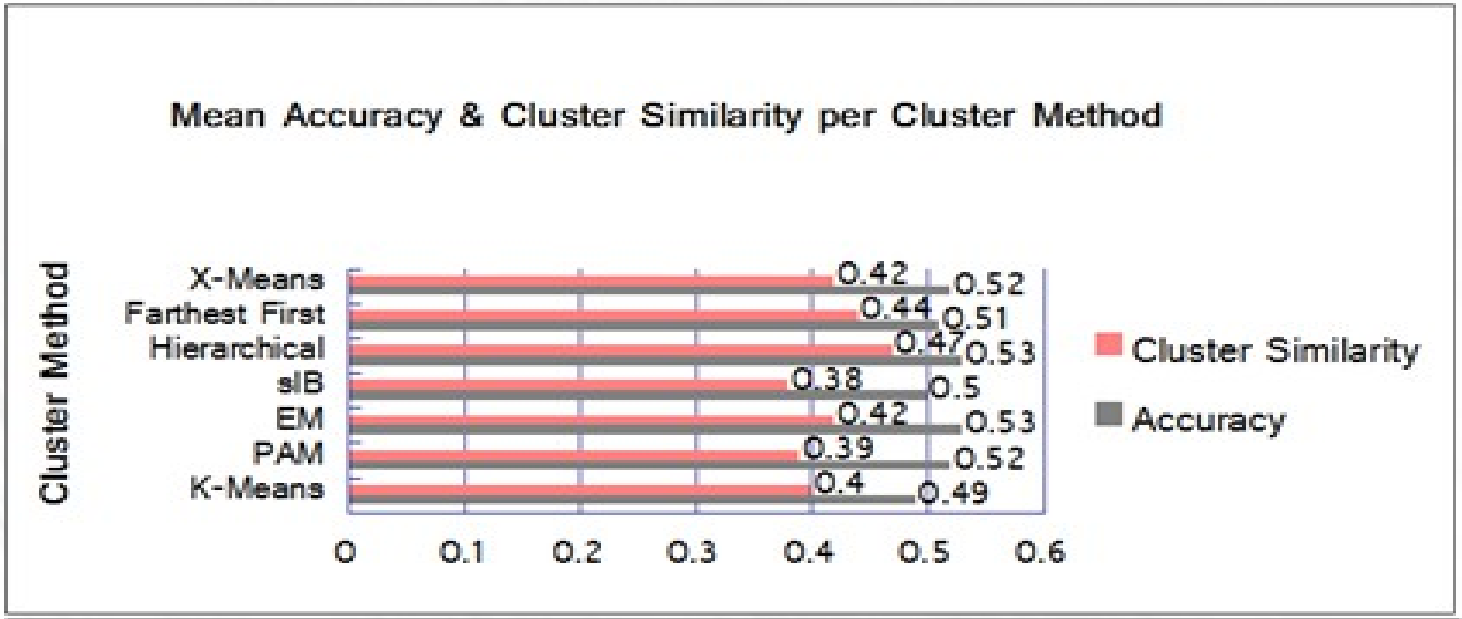
Cluster Evaluation Results (II)

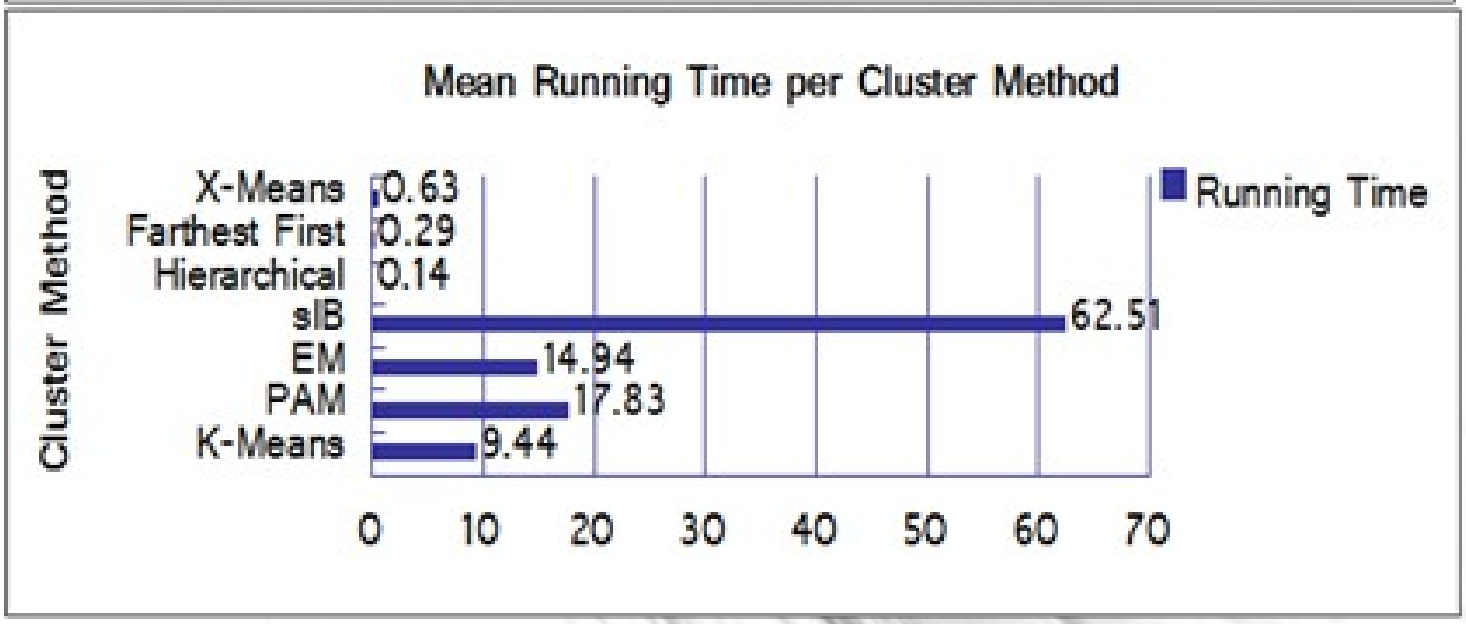
- Farthest First performed best in accuracy and cluster similarity
- EM performed worse in accuracy while Hierarchical performed worse in cluster similarity
- What if we cannot make the assumption: the datasets as “gold standard” ?

Experiment 2

- 6 datasets (randomly) with the provider 'The European Library'
- Clustering methods: X-Means, Farthest First, Hierarchical, sIB, EM, PAM, X-Means
- Cluster parameters
  - Transformation method for categorical data: Binaries
  - Selection of fields in the metadata
  - Distance measure: Euclidean, Jaccard, Manhattan
  - *k*: {2,3,4,5,6}

Cluster Evaluation Results





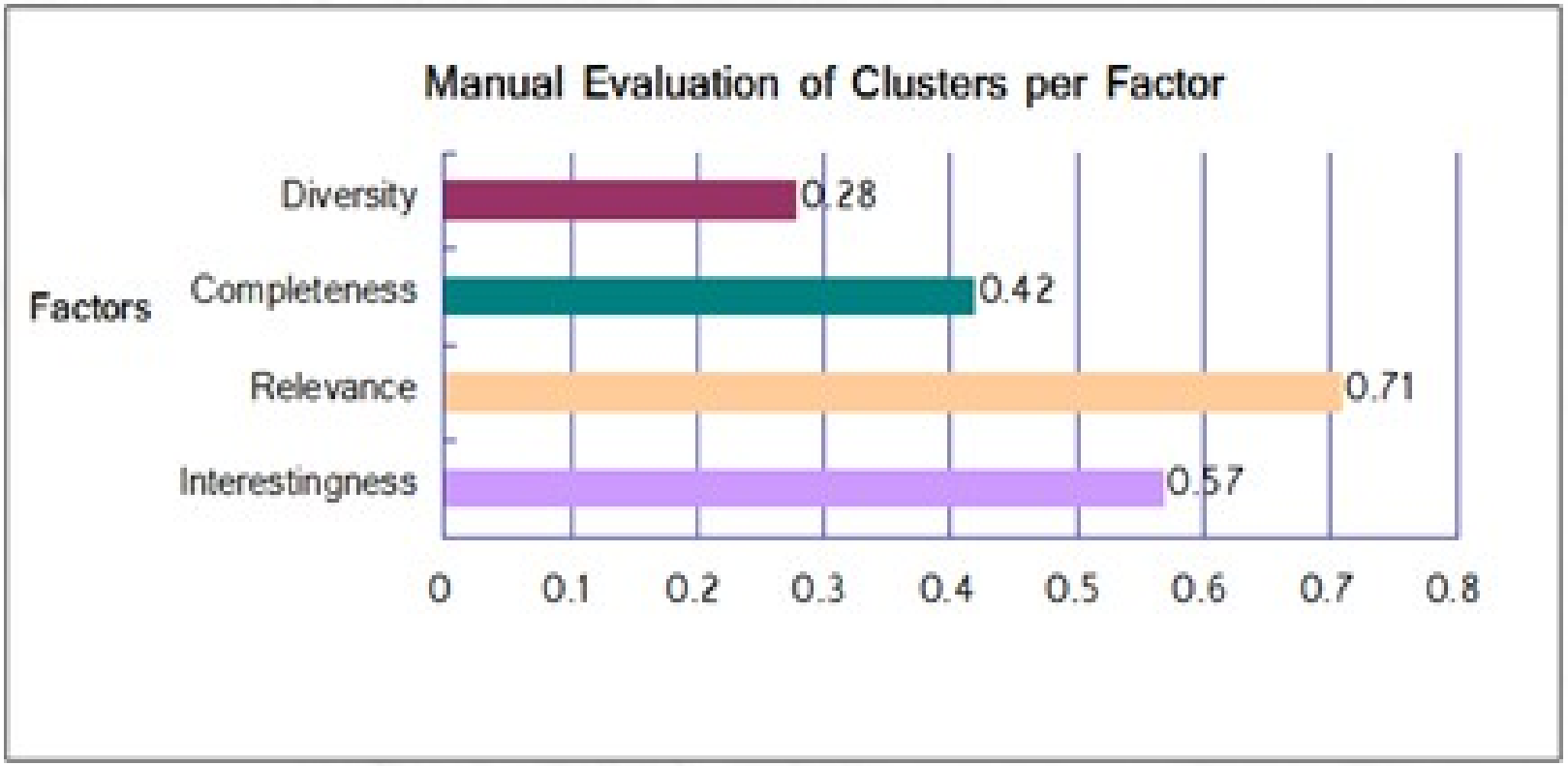
Cluster Evaluation Results (II)

- Hierarchical performed best among all the methods
- Best performance when k=2
- When k increases accuracy, cluster similarity decrease and running time increases

Experiment 3

- Selection of interesting datasets provided by domain experts at Europeana
- 9 datasets (randomly): 3 datasets from 3 aggregators
- Clustering method: Hierarchical
  - Distance: Euclidean
  - Threshold: 3.5
    - 7 clusters produced

Cluster Evaluation Results



Cluster Evaluation Results (II)

- Good scores for the evaluation factors except of completeness
- Difficult to know if good enough for Europeana only based on the manual evaluation results
- Evaluators judge clustering result itself, true clustering is not known
- We still do not know the correct number of clusters

Conclusion & Discussion

- The quality of clustering is dependent on the parameters we mentioned
- The algorithms showed better performance if k =2 but for an experiment of 100 datasets ?
- Hard to define the best parametric setting only based on a number of experiments
- If an algorithm discovers new structures this can be seen as a desirable result

Conclusion & Discussion (II)

- Good scores except of completeness for manual evaluation
- However, the evaluation is subjective & the domain experts judge the results itself
- We have shown a way to cluster Europeana objects which may be useful for Europeana

Future Work

- Adjust the hierarchical or other methods to determine optimum number of clusters
- Perform other experiments with big datasets to define the best clustering & setting
- Perform experiments with other clustering methods or by adjusting them

Questions

?