UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

*

N. Barbuti, T. Caldarola, S. Ferilli

A Graphic Matching Process for Searching and Retrieving Information in Digital Libraries of Manuscripts

IRCDL 2018
14th Italian Research Conference on Digital Libraries

Università di Udine
25-26 gennaio 2018

nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Humanistic databases

- Thesaurus Linguae Graece (http://stephanus.tlg.uci.edu)
- Integrated Archaeological Database (http://www.iadb.org.uk)
- World Digital Library (http://www.wdl.org/en)
- Musisque Deoque (www.mqdq.it/public)
- Trismegistos (www.trismegistos.org)
- Europeana (https://www.europeana.eu/portal/it)

Different levels of complexity

Scholars must have an hypotesis to confirm before searching database

nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it

Fourth Paradigm (Gordon Bell, 2012)

- Science informatics
- New approach to humanitistic database:
  - finding new research hypotheses from patterns inferred from large databases
  - discovering unexpected implications on consolidated research topics

nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Digital Recognition Systems

- Segmental approach: several prototypes based on HMMs
- Holistic approach: high percentage of recognized content, but processed image set not significant

Both approaches:

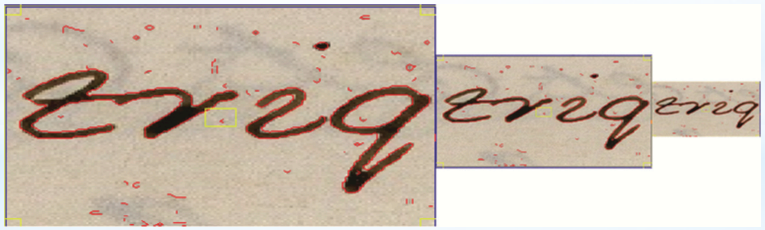
- Segmentation
- Data extraction (it's fine for printed contemporary documents)
- Matching with contents manually transcribed and/or annotated

nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it

Search algorithm

- It's based on the concept of "number of pyramid levels":
- The model consists of a number of graphic models having different resolution:
  - Image 1: 600x400 dpi
  - Image 2: 300x200 dpi, etc.

As a general rule, a good result is a ROI of width $2^{\text{Level Number}}-1$ (e.g. 8 pixels width allows one to use four pyramid levels)



nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DATA SCIENCE APPROACH

- Data Science perspective: “assumption-free” approach to querying several databases without any expectation on the results
- The system does not deal with a single, specific, pre-defined database, but it can be connected real time to several databases on-line
- The user can query all of them, or select one (or more than one) as relevant to his research objectives
- ‘Clusters’ of ‘similar’ poems might be identified in digital libraries or collections, or in literary corpora, and suggest new hypotheses on which an enquiry can be started using traditional approaches.
- interesting hypotheses may be investigated by analyzing both True Positive and False Positive outcomes

nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Experimental results – Test 1

- 3500 images belonging to 7 medioeval latin manuscripts, XI to XIII Century (Vatican Library) seemingly written by different amanuenses in many scriptoria
- Sample: “&”, “C”, “O”, “S”
- Parameter settings to create the shape model:
  - Deformation: 3
  - Resize: 40%
  - Minimum score: 60% / 80%

True Positives: 80% in ms 1, 2, 4 and 5; some False Positives in ms 5:

- Graph “C” = Graph “O”, Graph “S” (ancient style) = Graph “F”

The high percentage of homographs between these seemingly different manuscripts may suggest the use of a common writing canon along many centuries in distant scriptoria

nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Experimental results – Test 2

- Test case: ms A and B of Codex Sinaiticus (The British Library), generally considered different
- Shape model: graph “psi” in ms A
- Results:
  - True Positives: 50% in ms A, 25% in ms B
  - False Positives: 10% in ms A, 15% in ms B:
    - graph “phi” nearly homograph: 5% in ms A, 15% in ms B
    - graph “u” approximately homograph: 5% in ms B
  - Perhaps the two manuscripts were written by the same amanuensis

nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

CONCLUSION

- The paper describes a new approach to search and retrieve information in digital libraries based on the fourth paradigm of data science
- The training process of the ICRPad M-Evo module uses a matching algorithm based on contours shape recognition
- It uses the number of pyramid levels: the set of images processed is composed by models at different graphic resolutions
- The latest version of the module processes about 240.000 images/h with 60% / 90% of positive matches, depending on the parameter settings.

nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it

UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

*Thank You for attention

nicola.barbuti@uniba.it
t.caldarola@dabimus.com
stefano.ferilli@uniba.it