Similarity of mathematical texts

Radim Řehůřek, Petr Sojka

NLPlab, Masaryk University, Brno
Czech Republic

June 11, 2008

Goals, motivation

- DML-CZ project
- Finding similar articles
  - based on metadata (citations, fixed taxonomy – MSC)
  - based on fulltext similarity
- Automated MSC classification
  - what is MSC?
  - why automated?
- Plus incorporating results into DML workflow

Similarity search

- Goal is to offer similar articles to the user
  - based on same MSC, fulltext
- Fulltext
  - as an alternative to MSC, no manual labeling, metadata
  - with OCR, errors already at the character level; deep, fine analysis problematic
  - $\rightarrow$ simple (even stupid, but robust) IR techniques
- We used Vector Space Model with TFIDF and LSA

TFIDF vs LSA

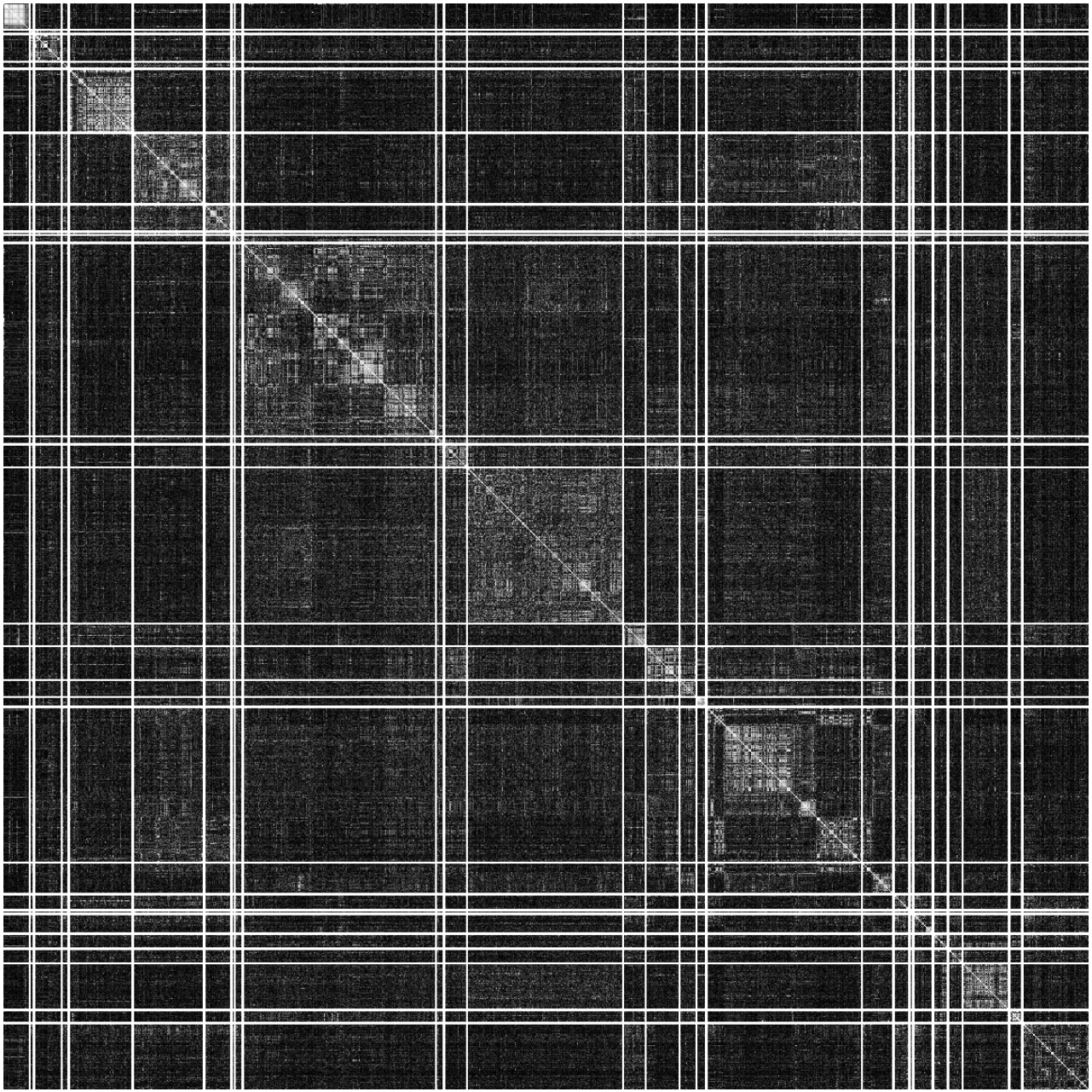
- both fully automated, language independent
- both look at term distribution
  - similarity metrics based on token (topic) overlap
- LSA purely statistical method of topic extraction
  - topic = linear combination of terms

Data

- fulltext taken from DML-CZ database, homogeneous wrt. source
- at the time of the experiments:
  - 4532 English articles
  - 595 Russian
  - 483 German
  - 276 French

Experiments, evaluation

- Experiments
  - pair-wise article similarity for TF*IDF, LSA
- Evaluation
  - comparison to metadata (MSC codes)
  - visual
  - lacking...
- UI: offer 10 most similar articles to the user

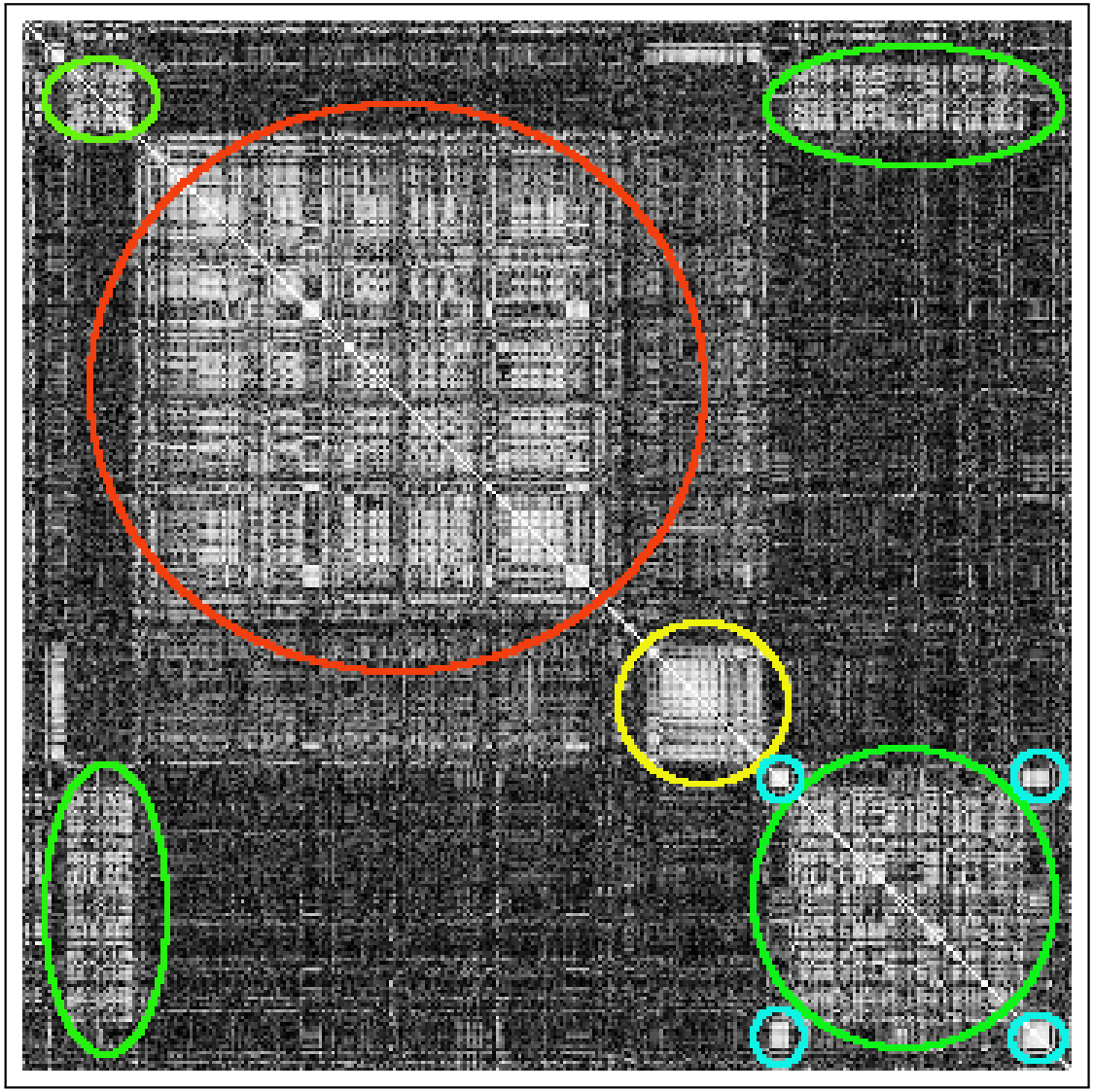


Category 20-xx: Group theory and generalizations

- 20-00 General reference works (handbooks, dictionaries, bibliographies, etc.)
- 20-01 Instructional exposition (textbooks, tutorial papers, etc.)
- 20-02 Research exposition (monographs, survey articles)
- 20-03 Historical (must also be assigned at least one classification number from Section 01)
- 20-04 Explicit machine computation and programs (not the theory of computation or programming)
- 20-06 Proceedings, conferences, collections, etc.
- 20Axx Foundations
- 20Bxx Permutation groups
- 20Cxx Representation theory of groups [See also 19A22 (for representation rings and Burnside rings)]
- 20Dxx Abstract finite groups
- 20Exx Structure and classification of infinite or finite groups
- 20Fxx Special aspects of infinite or finite groups
- 20Gxx Linear algebraic groups (classical groups) For arithmetic theory, see 11E57, 11H56; for geometric theory, see 14Lxx, 22Exx; for other methods in representation theory, see 15A30, 22E45, 22E46, 22E47, 22E50, 22E55
- 20Hxx Other groups of matrices [See also 15A30]
- 20Jxx Connections with homological algebra and category theory
- 20Kxx Abelian groups
- 20L05 Groupoids (i.e. small categories in which all morphisms are isomorphisms) For sets with a single binary operation, see 20N02; for topological groupoids, see 22A22, 58H05
- 20Mxx Semigroups
- 20Nxx Other generalizations of groups
- 20P05 Probabilistic methods in group theory [See also 60Bxx]

Category 20-xx: Group theory and generalizations

- 20.92 + 20.93, 20Mxx
- 20Dxx + 20Exx + 20Fxx
- 20.30, 20Kxx
- 20Lxx, 20Nxx



Future work

- better, real-world evaluation
- clickable UI :-)

Thank you for your attention