# *Document interlinking*
*in a*
# *Digital Mathematics Library*

Claude Goutorbe
(presented by Thierry Bouche)

Cellule MathDoc & Université Joseph Fourier, Grenoble

*Towards a Digital Mathematics Library*
Grand Bend, Ontario, 8-9 July 2009

## Outline

## The mathematical literature

### As a network

- Mathematical works build on previous results
- Any given work is part of a network of references
- This has always been the case, however
- New digital infrastructures can make this network explicit

## *Traversing the mathematical literature*

- Mathematicians enjoy good reference databases: Jahrbuch, Zentralblatt-MATH, Math Reviews
- They can be used as a linking hub
- From a given work's metadata to its reviews
- From the review to the actual text
- From a work to cited works
- From a work to citing works

## What is this about ?

- This work is not about extracting references from a digital file
- We assume that the digitization or publication process yields individual references
- These references are simple strings (no structure)
- We want to find database entries that describe the same work as the reference string

## Reference strings

- Typing errors, optical recognition errors
- Innacuracy (volume numbers, pages, publication year)
- Incomplete
- Translated titles
- and so on

# Reference strings

## *Example*

```
<bibitem>(V) L. Pontrjagin, Topological groups,
Princeton mathenatical séries, Vol. 2.,
(Princeton University Press), Princeton 1939.
</bibitem>
<bibitem>(VI) N. Bourbaki, Elements de mathématiques,
livres I, II, III, Actualités Scientifiques et Industrielles,
N° 848, 856, 916, 934, Paris, (Hermann) 1940-1942.
</bibitem>
<bibitem>(VII) A. Weil ;  L'intégration dans les groupes topologiques,
et ses applications. Actualités Scientifiques et Industrielles,
N° 869, Paris (Hermann) 1940.
</bibitem>
```

## Different techniques

- Field by field comparison
  - Involves some kind of parsing
  - This is notoriously difficult
  - Exact string comparisons cannot be used (different representations)
- Instead we do not try to identify subfields *a priori*

## String metrics

- Character based (Levenshtein distance and similar metrics)
  - Can handle typing and ocr errors on a local (field) basis
  - Do not work well on the complete reference string (different ordering of subfields)
- Token based
  - Match substrings (tokens) independently of their position
  - Using character $n$-grams as tokens allows for small mistakes and variations in spelling
  - Numerical tokens are of particular importance

## Numbers

```
Total number of journal articles: 413721

v = volume number
y = publication year
fp = first page number
lp = last page number
t = first (significant) title word

a = first author name (without initials)
Total number of different v|fp-lp strings: 376038 (90.89)
Total number of different t|fp-lp strings: 380623 (92.26)
Total number of different a|y|fp strings: 402594 (97.31)
Total number of different a|fp-lp strings: 406844 (98.33)
Total number of different a|v|fp strings: 410735 (99.28)
Total number of different a|y|fp-lp strings: 411959 (99.57)
Total number of different a|v|y|fp strings: 412350 (99.67)
Total number of different a|v|fp-lp strings: 412710 (99.76)
Total number of different a|v|y|fp-lp strings: 412889 (99.80)
```

- This shows that *numbers are important*

# Matching results

1. Journal articles from the Numdam project
   - Metadata is of good quality
   - All articles that are actually present in Zentralblatt-MATH are matched
2. Bibliographic references cited by these same articles
   - Metadata may be noisy because of optical recognition errors and inaccurate or incomplete because of authors' mistakes
   - Includes every possible kind of reference (journal articles, books, thesis, reports, ...).
   - The average rate of matches is 75 % of the total number of bibliographic items, and may grow up to 85 %, depending on the journal.
   - Results have been checked during the development of the software, meaning that these figures include a very low rate of irrelevant matches
3. Bibliographic references from the *Journal of Differential Geometry* (project Euclid)
   - Matching rate 89 %
   - No checking performed

## The matching strategy

- Generate an initial set of candidates using a boolean query
- Compute the cosine similarity for each candidate, using tri-grams
  - eliminate candidates whose score is below a certain threshold
- Check author names using approximate string matching
- Compare paging strings, if any
- Otherwise compute the Dice coefficient on number sets
- The process may stop at any point
- These steps use thresholds (empirically determined)
- It may be possible to use machine learning techniques to compute these values and build a decision tree
  - but building a good training set is not obvious

## Examples

```
[10] K. W. MORTON and S. SCHECHTER, On the stability of finite
 difference matrices (S.I.A.M. Series B, Vol. 1, 1965, pp. 119-128).

0.80021712845 Morton, K.W.; Schechter, S. On the stability of finite
 difference matrices. J. Soc. Ind. Appl. Math., Ser. B, Numer. Anal. 2, 119-128 (1965).

0.182050513664 Wright, Gretchen A foliated disk whose boundary
 is Morton's irreducible 4-braid. Math. Proc. Camb. Philos.
Soc. 128, No.1, 95-101 (2000).
```

- The initial query is for “morton”
- Only two items have at least 2 numbers in common with the reference string (1, 128)
- Candidate number 2 is eliminated (low cosine similarity)
- Candidate 1:
  - “morton” and “schechter” appear in the reference string
  - the paging string “119-128” is ok
  - together with a high cosine similarity (0.8), this is considered sufficient

## Examples continued

```
[2] N. V. BANITCHOUK,  V. M. PETROV,  F. L. TCHERNOUSSKO,
  Résolution numérique de problèmes aux limites variationnels
 par la méthode des variations locales.
 Journal de Calcul numérique et de Physique mathématique,
 tome 6, n. 6, Moscou, 1966, pages 947 Ã  961.

0.488957436865 Banichuk, N.V.; Petrov, V.M.; Chernous'ko, F.L.
 The solution of variational and boundary value problems by
 the method of local variations.
 U.S.S.R. Comput. Math. Math. Phys. 6, No.6, 1-21 (1966);
 translation from Zh. Vychisl. Mat. Mat. Fiz. 6, 947-961 (1966).
```

- Dice coefficient of $2.\frac{4}{14} = 0.57$, and a cosine similarity of $0.48$.

## When numbers cannot be used

- "Not enough" numbers in the reference string
- Too many wrong numbers
- Try to match the usual three parts: authors, title, bibliographic data
- Authors and title are matched using approximate substring matching
    - may fail

```
[5] P G CIARLET,  The Finite Element Method for Elliptic Equations,
 North-Holland, Amsterdam, 1978
```

should map to

```
Zbl 0383.65058 Ciarlet, Philippe G.
The finite element method for elliptic problems.
Studies in Mathematics and its Applications. Vol. 4. Amsterdam - New
York - Oxford: North-Holland Publishing Company (1978)
```

- This is a common case of failure
- A token based metric might be more appropriate
- Needs to be explored further

# When numbers cannot be used, continued

- Comparing journal titles
  - compute Dice coefficient using common prefixes

    ```
    J. Diff. Geom.
    Jour. of Differ. Geometry
    ```
  - common prefixes: "J", "Diff" and "Geom", giving a Dice coefficient of $2.\frac{3}{3+3} = 1$
- Book data is a challenge of its own
  - data in the database is very complete (publisher, publication place, edition statement, collection, etc...)
  - whereas authors usually give a small subset of these fields
  - multiple edition and reprints are not easy to distinguish
  - publication years are surprisingly often different
- Compute the set of common tokens (words)
- Give more weight to less frequent terms (inverse database frequency)
- Seems to work well in practice
- Needs further investigation !

# Conclusion

- The software works very well on article references (where several numbers are present)
- It has some difficulties on books, when a number of different editions or reprints exist in the database
- Improving the matching rate in this case probably requires a deeper analysis of the input string and/or using different metrics