

Topic Similarity in Information Retrieval

Examples and Experience of NLP Centre and LEMMA Projects

Petr Sojka

Laboratory of Electronic and Multimedia Applications and
Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

`sojka@fi.muni.cz`

PV211 Intro to Information Retrieval: LDA

Coping with Information Overload by Filtering of *Big Data*



Life is searching: group **similar** and narrow focus of search in [your] Big Data.

Similarity types: from **plagiarism** (similarity on n -grams, narrative similarity, evolved into <http://theses.cz>) to **thematic, topical similarity**.

Levels of content processing: strings → words and collocations → semantics (word meaning) → information (knowledge).

Grabbing the essence (content) of documents: **topical modelling**.

Zde zadejte hledané slovo nebo slova oddělená čárkou (= operátor ACCRUE) nebo jiným operátorem (lze použít tlačítek Volba operátorů). Slova budou hledána ve všech přípustných tvarech (kromě změny kmenové souhlásky) bez zřetele na velká/malá písmena. Při hledání fráze (např. univerzita karlova) se tato dvě slova zadají neoddělená čárkou. Při hledání slova v přesném tvaru (bez skloňování/časování) se slovo uvede v uvozovkách, např. „VŠE“.

Hledaný text

- Ottova encyklopedie obecných vědomostí® Ottova encyklopedie nové doby
 hledat v plných textech hesel pouze v názvech hesel volný text

Vazby mezi výrazy: ACCRUE (čím více, tím lépe) AND (a) OR (nebo) NOT (ne)

Topical Similarity in Digital Mathematics Library

Similar articles to article

[CHEN, HUANYIN](#)








Strong separativity over exchange rings. (English). Czechoslovak Mathematical Journal, vol. 58 (2008), issue 2, pp. 417-428

[-> Back to article](#)

Method LSI

Generalized $\mathcal{V}\mathcal{S}$-rings ...	
Exchange rings in whic...	
Rings which have proje...	
Epimorphisms of regul...	
Von Neumann regular ri...	
A general theory of Fo...	
On $\mathcal{V}\mathcal{S}$-rings and unit...	
Extensions of $\mathcal{S}\mathcal{G}\mathcal{M}\mathcal{S}$-rings	
$\mathcal{S}\mathcal{E}\mathcal{S}$-rings and differen...	

Method RP

Exchange rings with st...	
Generalized $\mathcal{V}\mathcal{S}$-rings ...	
Exchange rings in whic...	
Rings which have proje...	
$\mathcal{S}\omega$-$\mathcal{I}\mathcal{S}$-generated u...	
Diagonal reductions of...	
Von Neumann regular ri...	
Steady ideals and rings	
Dualities over compact...	
The p. p. ring and the...	

Method TFIDF

Exchange rings with st...	
Exchange rings in whic...	
Diagonal reductions of...	
The least separative c...	
Note on the congruence...	
Extension of measure-l...	
Integration in partial...	
Modularity and distrib...	
On abelian groups by w...	
Extensions of $\mathcal{S}\mathcal{G}\mathcal{M}\mathcal{S}$-rings	

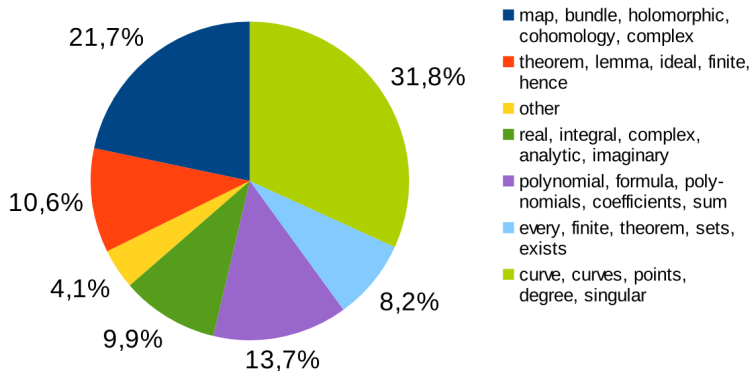
We appreciate your feedback to the methods which determine similarity of articles (e.g. which method is better, ...).
Please [contact us](#). It will be helpful for future development.

[-> Back to article](#)

- ▶ 2005, GVP, Radim Řehůřek and Jan Pomikálek
- ▶ 2006, gensim, different machine learning methods as Random Projections, TFIDF word weighting, Latent Semantic Indexing/Analysis, Latent Dirichlet Allocation
- ▶ 50,000+ full-texts on <http://dm1.cz>

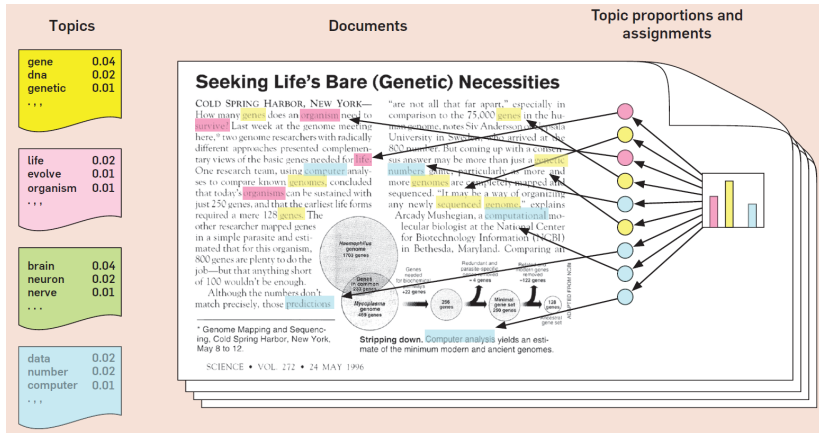
Leading Edge Example: Automated Meaning Picking from Texts

LDA Topics Pie Chart for math.0406240



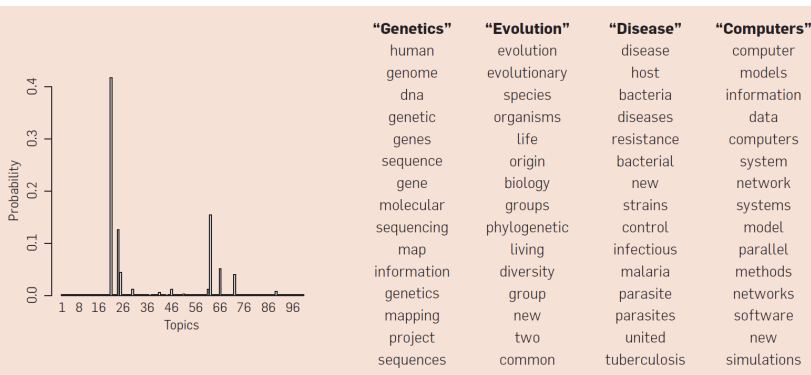
Probabilistic Topical Modelling: Latent Dirichlet Allocation

- ▶ topic: weighted list of words
- ▶ document: weighted list of topics



Topical Modelling: Latent Dirichlet Allocation II

- ▶ all topics computed automatically from document corpora



Content Similarity Results in EuDML

Within *European Digital Mathematics Library*, *EuDML*, project EU CIP-ICT-PSP we have developed and delivered technology for **similarity** (gensim), document **conversions** (Braille) and **accessibility** (math OCR), NLP content **normalization** (Mathml2text).

The screenshot shows the EuDML website interface. At the top left is the logo "EuDML" and "The EUROPEAN DIGITAL MATHEMATICS LIBRARY". On the top right, there is a language dropdown menu set to "English (en)", a user name "Jane Doe", and a "Log Out" link. Below this is a search bar with the placeholder text "Title, Author, Keyword, Citation, Date..." and a "Search" button. A dark navigation bar contains links for "Home", "Advanced Search", "Browse by Subject", "Browse by Journals", and "Refs Lookup". The main content area displays a search result for a document titled "On the solution of the differential equation $f(x, y, y^{(1)}, \dots, y^{(n)}) = 0$ ". The author is listed as "Smbat Abian, Arthur B. Brown (1958)" and the journal as "Bollettino dell'Unione Matematica Italiana". A "Similarity:" label is followed by a progress bar that is approximately 75% full. Below this, another result is shown for "Superposition of imbeddings and Fefferman's inequality" by "Miroslav Krbeč, Thomas Schott (1999)", also from "Bollettino dell'Unione Matematica Italiana", with a "Similarity:" label and a progress bar that is approximately 50% full.

Demo of math search in EuDML

[Help About](#)



How to write query

```
<math>\langle \mrow \langle \msup \langle \mi \rangle x \langle / \mi \rangle \langle \mn \rangle 2 \langle / \mn \rangle \langle / \msup \rangle \langle \mo \rangle + \langle / \mo \rangle \langle \msup \rangle \langle \mi \rangle y \langle / \mi \rangle \langle \mn \rangle 2 \langle / \mn \rangle \langle / \msup \rangle \langle / \mrow \rangle \langle / \math \rangle</pre>
```

Canonicalized MathML query:

```
<math xmlns="http://www.w3.org/1998/Math/MathML">
  <mrow>
    <msup> <mi>x</mi><mn>2</mn></msup>
    <mo>+</mo>
    <msup> <mi>y</mi><mn>2</mn></msup>
  </mrow>
</math>
```

Search in:

Total hits: 36817, showing 1-30. Searching time: 116 ms

Finite Precision Measurement Nullifies Euclid's Postulates

... and the unit circle $x^2 + y^2 = 1$ are both dense but they do not intersect, in contradiction to Euclid's postulates ...

score = 3.2980976

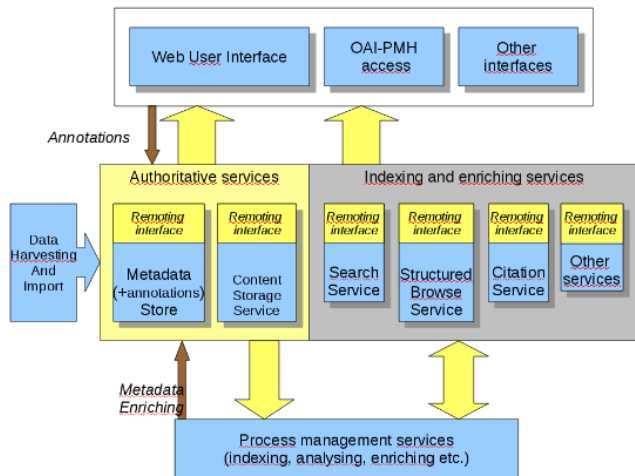
arxiv.org/abs/quant-ph/0310035 - cached XHTML

COMMENT ON RECENT TUNNELING MEASUREMENTS ON Bi22Sr2CaCu2O88

... gap, (b) s-wave gap, and (c) $s_x^2 + y^2$ gap.

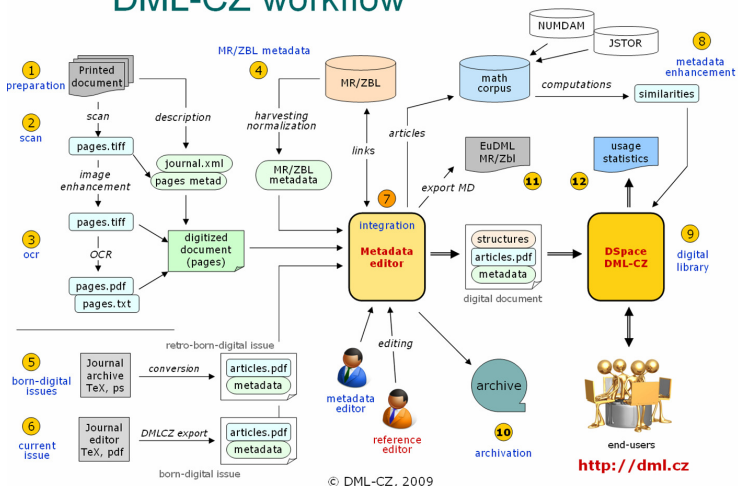
score = 4.0000000

Document engineering and workflows including [Math] OCR.



Digital Library Service Architecture and Workflow (DML-CZ)

DML-CZ workflow



Data Visualization and Representation

Došlý title author

Zobrazení

Přiblížit

Proceedings of EQUADIFF 10, Pr...
half-linear equation
Došlý, Ondřej
growth, boundedness
qualitative theory of half-lin...
variational method
Poincaré's identity
Asymptotic behaviour of oscill...
On an asymptotic behaviour of...
exponential stability and expo...
Comparison theorems for nonlin...
Substitution method for g...
A Remark on the Oscillatorines...
Asymptotic properties
On existence of Khneser solutio...
Asymptotic behaviour of the so...

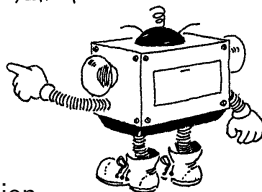
Asymptotic behaviour of osc...
language: eng
title: Asymptotic behaviour of oscillatory solutions of a fourth-order nonlinear differential equation@en
summary: Asymptotic behaviour of oscillatory solutions of the fourth-order nonlinear differential equation with quasiderivates $y^{[4]} + r(t)f(y) = 0$ is studied.@en

Ontologie: /22-rdf-syntax-ns# Perspektiva: DML -12

- ▶ Semantic similarity indexing and search of big (continuous stream of) data. Client (search) and server (indexing) architecture.
- ▶ Developed by NLP lab PG student Radim Řehůřek (awarded in Česká hlava competition in 2011).
- ▶ Leading edge machine learning methods implemented.
- ▶ Used in 60+ local, EU or worldwide projects, 260+ citations.
- ▶ Typical deployment and fine-tuning scenario: expressing data as words (features) → configuration of topic modelling of features → setting of gensim methods and tuning parameters → usage in an application with proper visualization interface.

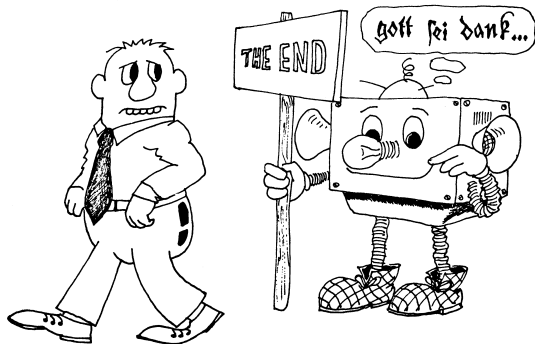
- ▶ most work done by students themselves with agile techniques, XP

Conclusions and Mutual Research Interests



- ▶ similarity by topical modelling, document filtering and visualization
- ▶ semantic, meaning computations and modelling of natural language texts (natural NLP)
- ▶ personal research interests: random walking for disambiguation, math (tree) indexing and similarity

Yes, we can!



Credits: Jiří Franek (illustrations)

- ▶ NLP Centre: <http://nlp.fi.muni.cz/>
- ▶ Topical modelling: <https://mir.fi.muni.cz/gensim/>
- ▶ Math Information Retrieval: <https://mir.fi.muni.cz>
- ▶ DML-CZ project: <http://dml.cz>, <http://project.dml.cz>
- ▶ EuDML project: <http://eudml.cz>,
<http://project.eudml.cz>
- ▶ LEMMA: <http://www.fi.muni.cz/lemma/>