

Topic Similarity in Information Retrieval

Examples and Experience of NLP Centre and LEMMA Projects

Petr Sojka

Laboratory of Electronic and Multimedia Applications¹ and
Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

sojka@fi.muni.cz

Seznam day, April 16th, 2014

¹Donor of today's catering :-)

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**.

Prehistoric Example: Project Ottův Slovník naučný, 1998

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

Grabing 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 encyklopédie obecných vědomostí® Ottova encyklopédie nové doby
 hledat v plních 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 \\$VS\\$-rings ...](#)
- [Exch NUMDAM: Generalized \\$VS\\$-rings and von Neumann regular rings](#)
- [Exchange rings in whic...](#)
- [Rings which have proje...](#)
- [Epimorphisms of regula...](#)
- [Von Neumann regular ri...](#)
- [A general theory of Fo...](#)
- [On \\$VS\\$-rings and unit...](#)
- [Extensions of \\$GMS\\$-ring](#)
- [\\$ES\\$-rings and differen...](#)

Method RP

- [Exchange rings with st...](#)
- [Generalized \\$VS\\$-rings ...](#)
- [Exchange rings in whic...](#)
- [Rings which have proje...](#)
- [\\$Omega_1\\$-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 \\$GMS\\$-ring](#)

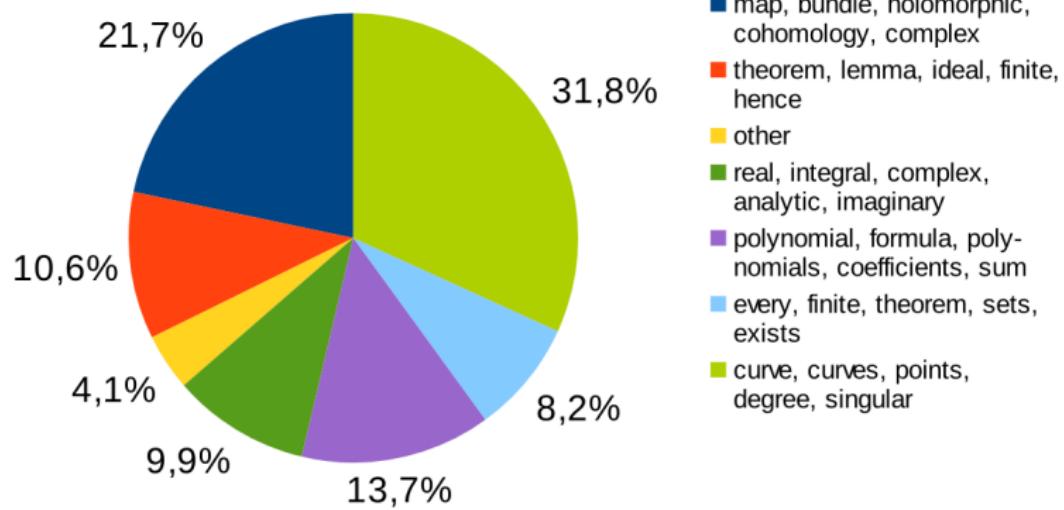
We appreciate your feedback to the methods which determine similarity of articles (e.g. which method is better, ...). Please [contact us](#). It will be helpful for future development.

> Back to article

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

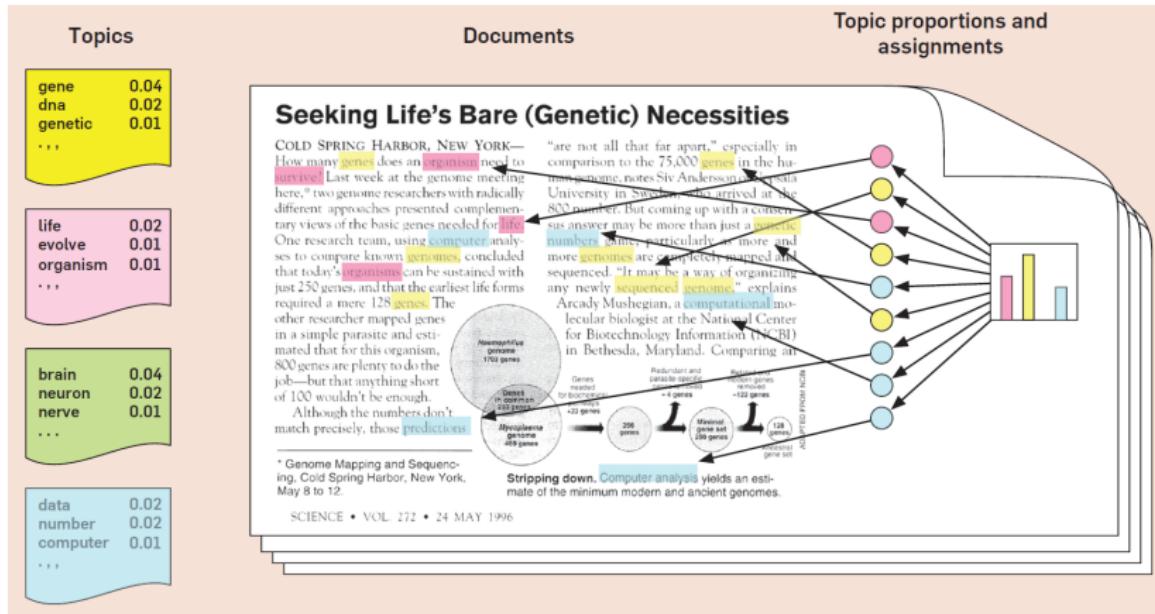
Leading Edge Example: Automated Meaning Picking from Texts

LDA Topics Pie Chart for math.0406240



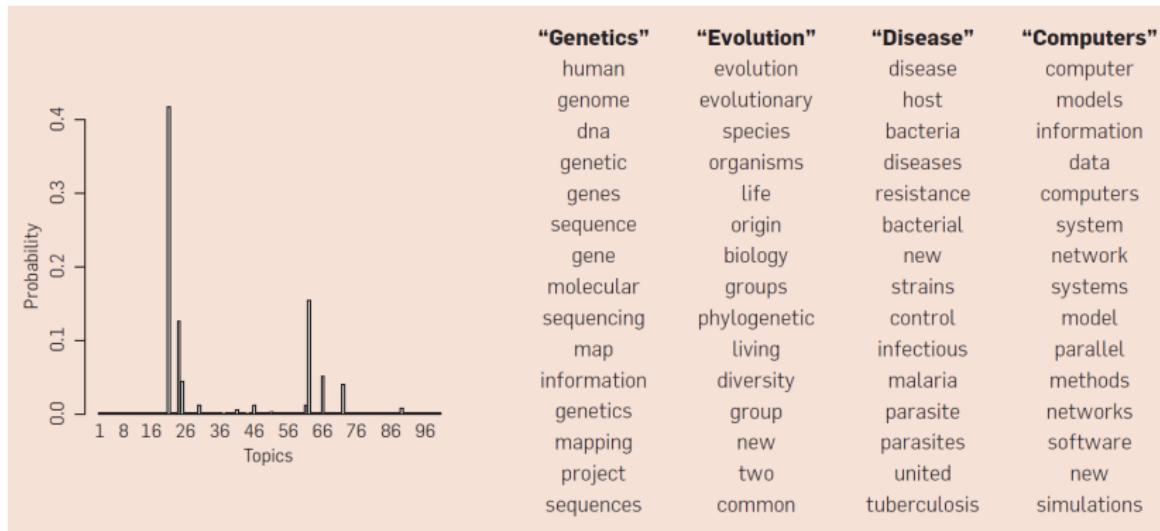
Probabilistic Topical Modelling: Latent Dirichlet Allocation

- ▶ 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 homepage with a search interface. The search bar contains "Title, Author, Keyword, Citation, Date...". Below it are buttons for "Search" and "Advanced Search". A navigation menu includes "Home", "Advanced Search", "Browse by Subject", "Browse by Journals", and "Refs Lookup". A sidebar displays search results for documents related to "On oscillation criteria for third order nonlinear delay differential equations". The first result is "On the solution of the differential equation $f(x, y, y^{(1)}, \dots, y^{(n)}) = 0$ " by Simbat Abian and Arthur B. Brown (1958). The second result is "Superposition of imbeddings and Fefferman's inequality" by Miroslav Krbec and Thomas Schott (1999). Both results show a similarity score represented by a progress bar.

EuDML | The EUROPEAN DIGITAL MATHEMATICS LIBRARY

English (en) ▾ Jane Doe Log Out

Title, Author, Keyword, Citation, Date... Search

Home Advanced Search Browse by Subject Browse by Journals Refs Lookup

Displaying similar documents to “On oscillation criteria for third order nonlinear delay differential equations”

On the solution of the differential equation $f(x, y, y^{(1)}, \dots, y^{(n)}) = 0$

Simbat Abian, Arthur B. Brown (1958)
Bollettino dell'Unione Matematica Italiana
Similarity:

Superposition of imbeddings and Fefferman's inequality

Miroslav Krbec, Thomas Schott (1999)
Bollettino dell'Unione Matematica Italiana
Similarity:

Math Search Interface EuDML

Demo of math search in EuDML

[Help](#) [About](#)



How to write query

```
<math><mrow><msup><m>x</m></msup><mn>2</mn></mrow><mo>+</mo><msup><m>y</m></msup><mn>2</mn></mrow></math>
```

[cl](#)

Canonicalized MathML query:

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

[cl](#)

Search In: MREC 2011.4.439 ▾ [Search](#)

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

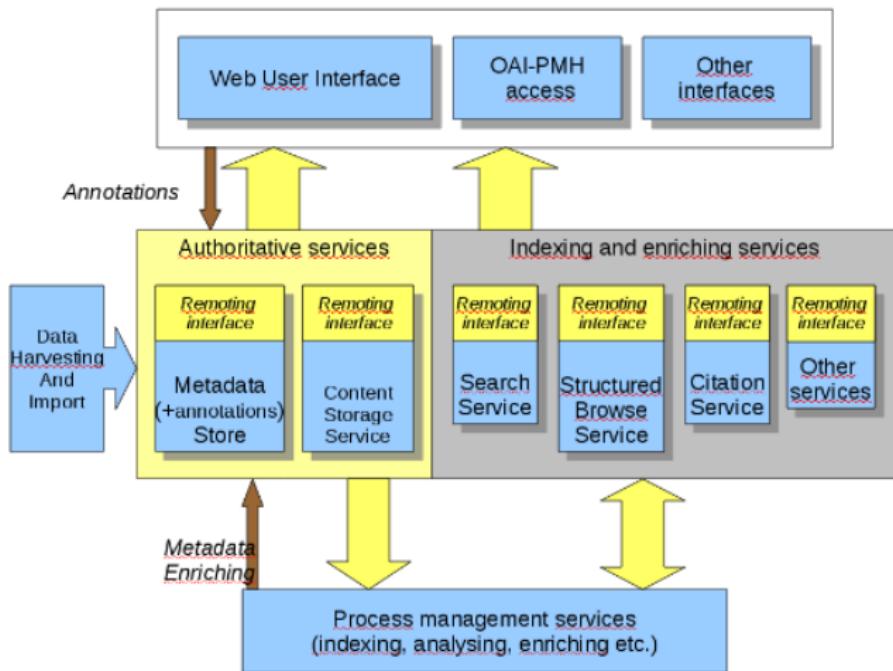
COMMENT ON RECENT TUNNELING MEASUREMENTS ON Bi22Sr22CaCu22O88

... gap, (b) s-wave gap, and (c) $s_{x^2+y^2}$ gap.

[arxiv.org/abs/1302.4006](#)

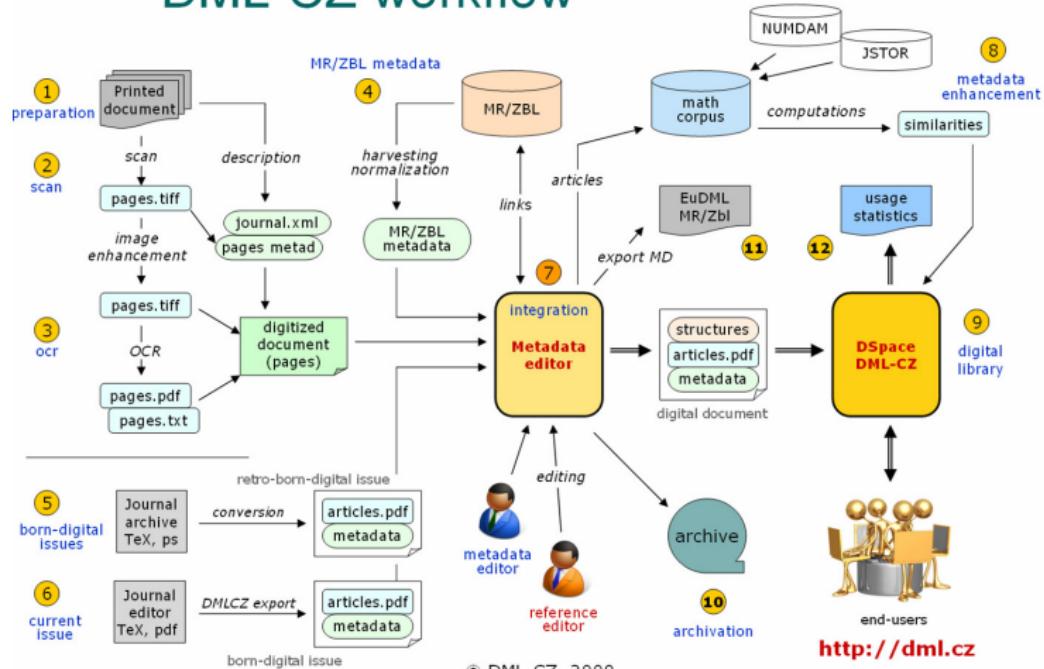
Digital Library Service Architecture and Workflow (EuDML)

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...

Došlý, Ondřej

half-linear equation

qualitative theory of half-lin...

Asymptotic behaviour of oscill...

On an asymptotic behaviour of ...

Exponential stability and expo...

A Remark on the Oscillatorines...

Comparison theorems for nonlin...

Substitution method for g...

On existence of Kneser solution...

Asymptotic properties

Asymptotic behaviour of the so...

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 quasiderivatives $y^{[4]} + r(t)f(y) = 0$ is studied.@en

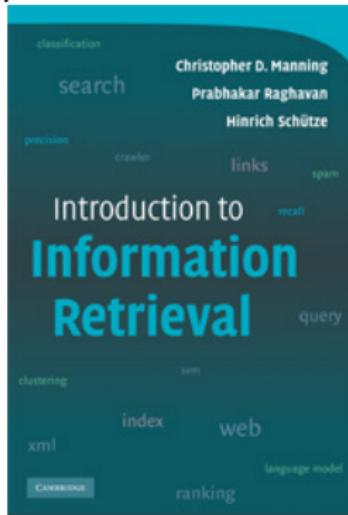
Ontologie: /22-rdf-syntax-ns# Perspektiva: DML -12

Award Winning Topic Similarity Framework **gensim**

- ▶ Semantic similarity indexing and search of big (continuous stream of) data. Client (search) and server (indexing) architecture.
- ▶ Developed by NLPlab PG student Radim Řehůřek (awarded in Česká hlava competition in 2011).
- ▶ Leading edge machine learning methods implemented.
- ▶ Used in 50+ local, EU or worldwide projects, 88+ 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.

Teaching Laboratory build with Constructivism Principles

- ▶ new course PV211, *Úvod do získávání informací, Introduction to Information Retrieval*



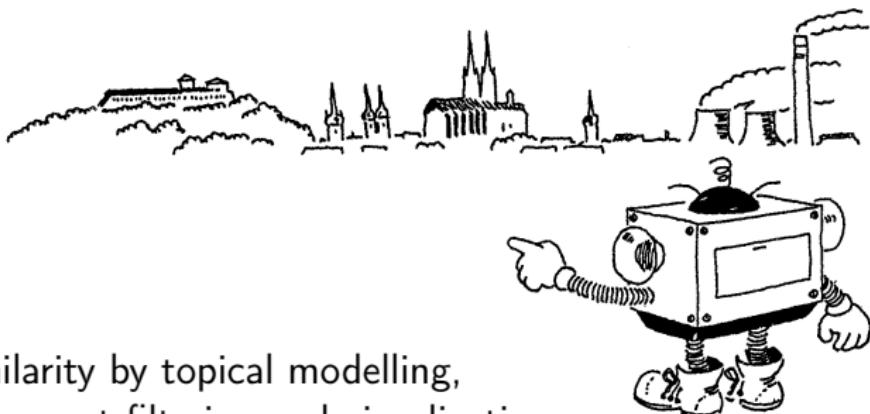
- ▶ most work done by students themselves with agile techniques, XP

New course PV211: Introduction to Information Retrieval

- ▶ In Spring 2014 for the first time: 100 students registered, 60 enrolled
- ▶ Invited lecture by Seznam (Roman Rožník)²
- ▶ students motivated by Khan Academy movies, premium tasks,...
- ▶ further cooperation, continuation course?

²cca 10 years ago lectured in Brno Štěpán Škrob and Ivo Lukaševič flied from Prague to brainhunt our students

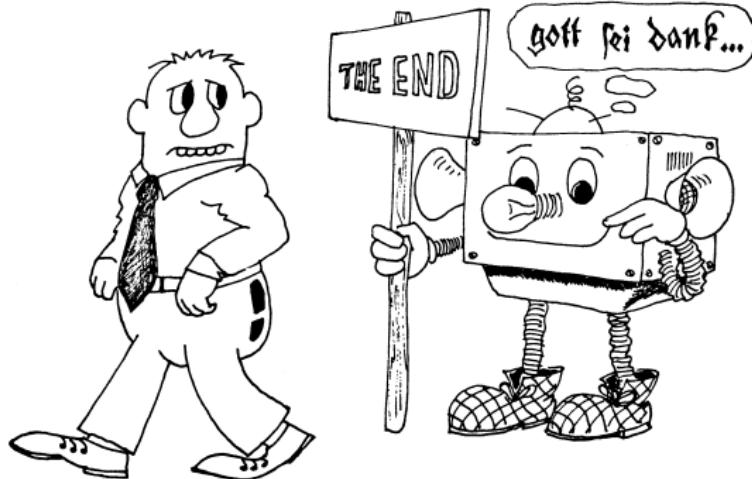
Conclusions and Mutual Research Interests



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

That's it!

Yes, we can!



Credits: Jiří Franek (illustrations)

Links

- ▶ 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/>