TRENDS IN EDUCATIONAL RESEARCH:

SEMANTIC ANALYSIS OF PAPER ABSTRACTS

Stefan Trausan-Matu*, Philippe Dessus**,
Cristian Ionut Paraschiv*, & Mihai Dascalu*

*RACAI, "Politehnica" Univ. of Bucharest, Romania

**LSE, Univ. Grenoble Alpes, France

CC BY NC SA

Outline

1. Considerations
2. Experiment
3. Further work

Refs available at

http://www.citeulike.org/user/pdessus/tag/educmap

1. CONSIDERATIONS

Pros & Cons of Co-citation/biblio. Coupling-based Analyses

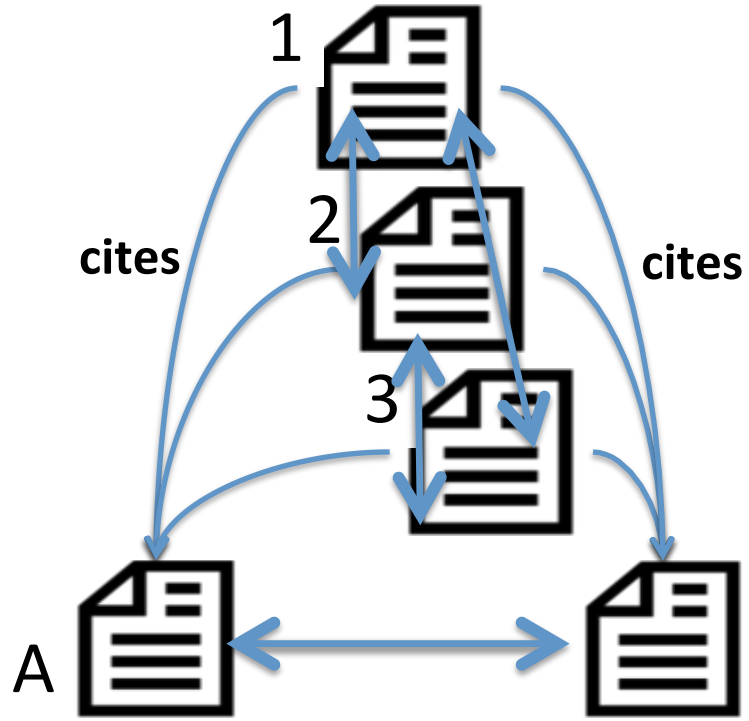
vs.

Text-based Analysis

[4]Co-citation Analysis, Biblio. Coupling and Semantic Comparison

- How to measure the relatedness between papers?

Bet. A & B: co-citation analysis
Bet. 1-3: biblio. coupling analysis



http://en.wikipedia.org/wiki/Co-citation

Semantic comparison of papers' abstracts

5 Co-citation/Biblio. Coupling Analyses

Pros	Cons
Co-occurrence-based and bibliography-centered, thus fast to process with very large sets of papers	Citation strategies are influenced by researchers' practices [Sword 12] and motivations [Ding et al. 14] - Citation number differ across domains - Influenced by authors' impact factor increasing endeavors - "Homage" effect may entail overcitation (a.k.a.,"preferential attachment")
Bib. coupling captures collaboration & communities =>network extraction easier	Assume all references are of similar value [Ding et al. 14], e.g., don't address the number of times a ref. is cited in a paper [Wan & Liu 14], or how it is cited (where in the paper) Two citations of a given paper may refer to different aspects of the cited paper

Text-based Approach

Pros	Cons
Content-based, thus more accurate than Co-Citation Analysis to assess paper's similarity, at least to measure cohesion, topic similarity	Semantic analysis is more computationally demanding: fewer data is computable in one pass
Reduced variability of summaries content and size across domains	Network extraction is more difficult; no obvious relationship with impact factor measures
Co-cit.: Texts already there (citations come over time)	Semantic similarity bet. papers doesn't necessarily mean authors' "community of thought" , rather replicating, copying

2. EXPERIMENT

Use the semantic analysis (LSA/LDA/ WordNet) of paper summaries to compare them directly (instead of using references as a proxy).

LSA Example [Landauer & Dumais 97]

§1: *Human* machine *interface* for ABC *computer* applications
§2: A *survey* of *user* opinion of *computer system response time*
§3: The *EPS* *user interface* management *system*
§4: *System* and *human* *system* engineering testing of *EPS*
§5: Relation of *user* perceived *response time* to error measurement
§6: The generation of random, binary, ordered *trees*
§7: The intersection *graph* of paths in *trees*
§8: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
§9: *Graph minors*: A *survey*

H-M Interface

Graphs

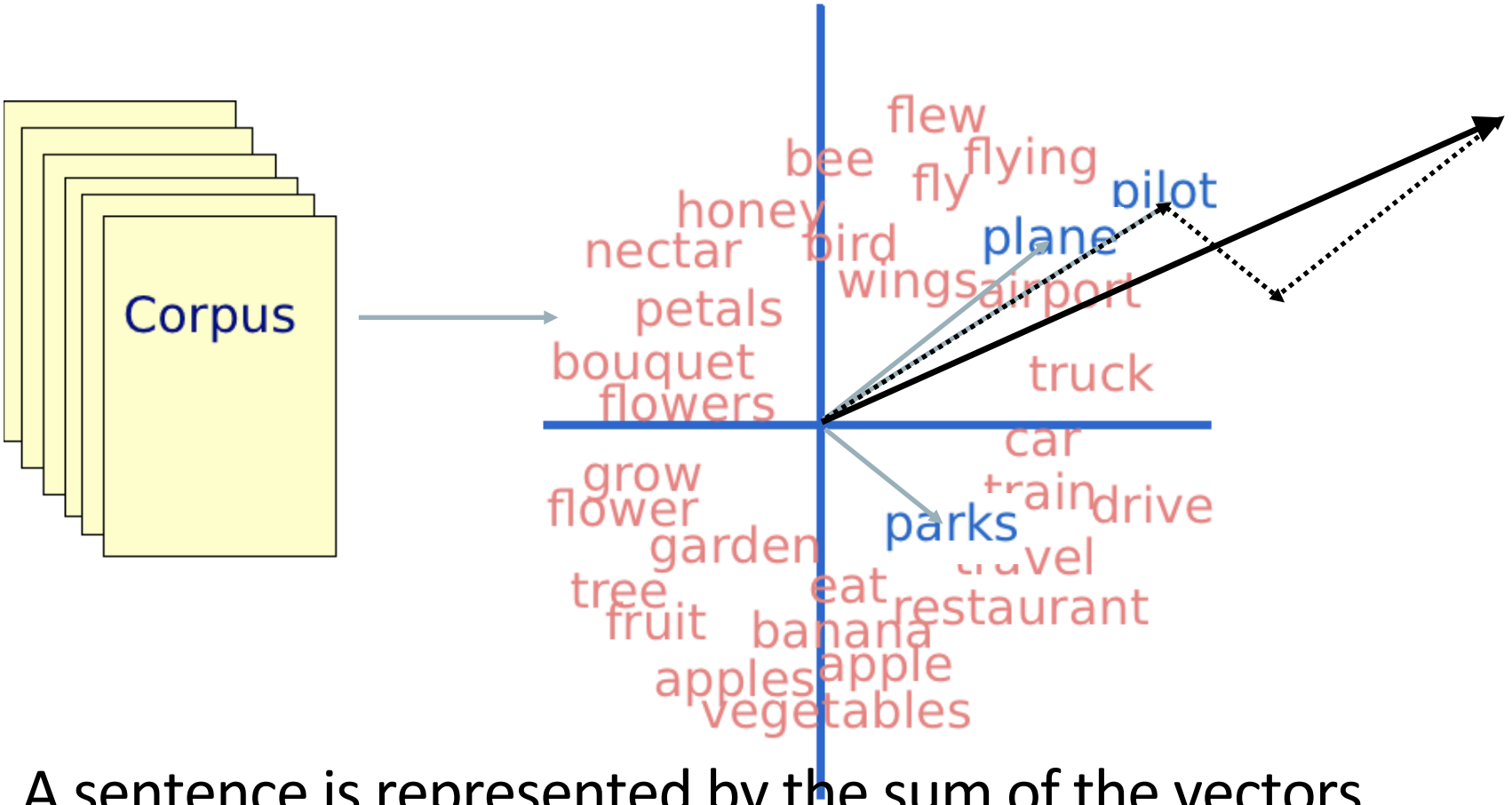
Human	1	0	0	1	0	0	0	0	0	.16	.40	.38	.47	.18	-.05	-.12	-.16	-.09
interface	1	0	1	0	0	0	0	0	0	.14	.37	.33	.40	.16	-.03	-.07	-.10	-.04
computer	1	1	0	0	0	0	0	0	0	.15	.51	.36	.41	.24	.02	.06	.09	.12
user	0	1	1	0	1	0	0	0	0	.26	.84	.61	.70	.39	.03	.08	.12	.19
system	0	1	1	2	0	0	0	0	0	.45	1.23	1.05	1.27	.56	-.07	-.15	-.21	-.05
response	0	1	0	0	1	0	0	0	0	.16	.58	.38	.42	.28	.06	.13	.19	.22
time	0	1	0	0	1	0	0	0	0	.16	.58	.38	.42	.28	.06	.13	.19	.22
EPS	0	0	1	1	0	0	0	0	0	.22	.55	.51	.63	.24	-.07	-.14	-.20	-.11
survey	0	1	0	0	0	0	0	0	1	.10	.53	.23	.21	.27	.14	.31	.44	.42
trees	0	0	0	0	0	1	1	1	0	-.06	.23	-.14	-.27	.14	.24	.55	.77	.66
graph	0	0	0	0	0	0	1	1	1	-.06	.34	-.15	-.30	.20	.31	.69	.98	.85
minors	0	0	0	0	0	0	0	1	1	-.04	.25	-.10	-.21	.15	.22	.50	.71	.62

r(human,user)= -.38

r(human,user)=.94

Latent Semantic Analysis: Sentence/Document Representation [Landauer & Dumais 90]

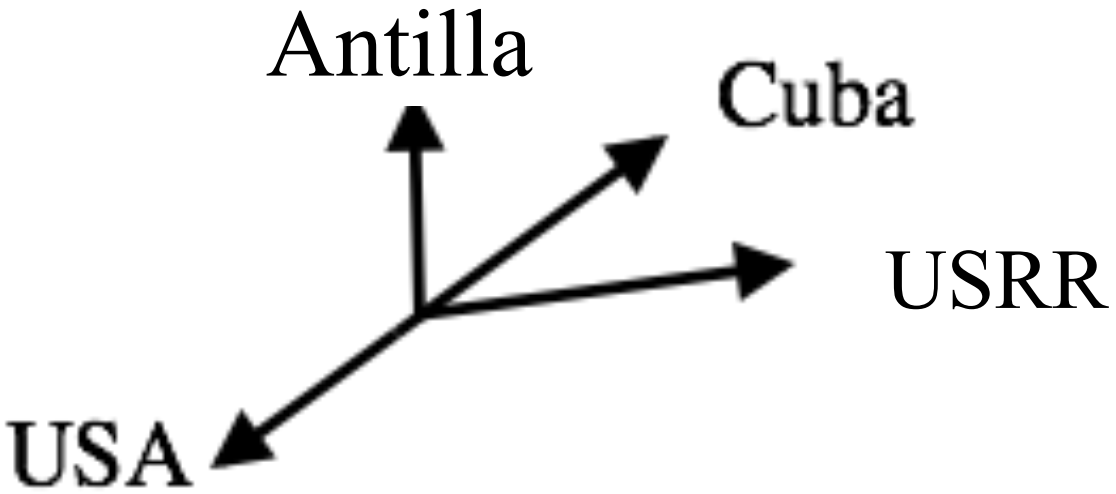
"The pilot parks the plane"



A sentence is represented by the sum of the vectors of its words [Lemaire & Denhière '05]

LSA Captures some (but not all!) Complex Word Semantic Relationships

Hypothetical Vectors [After Tversky 77]



Topic Modelling with Latent Dirichlet Allocation [Blei 2012]

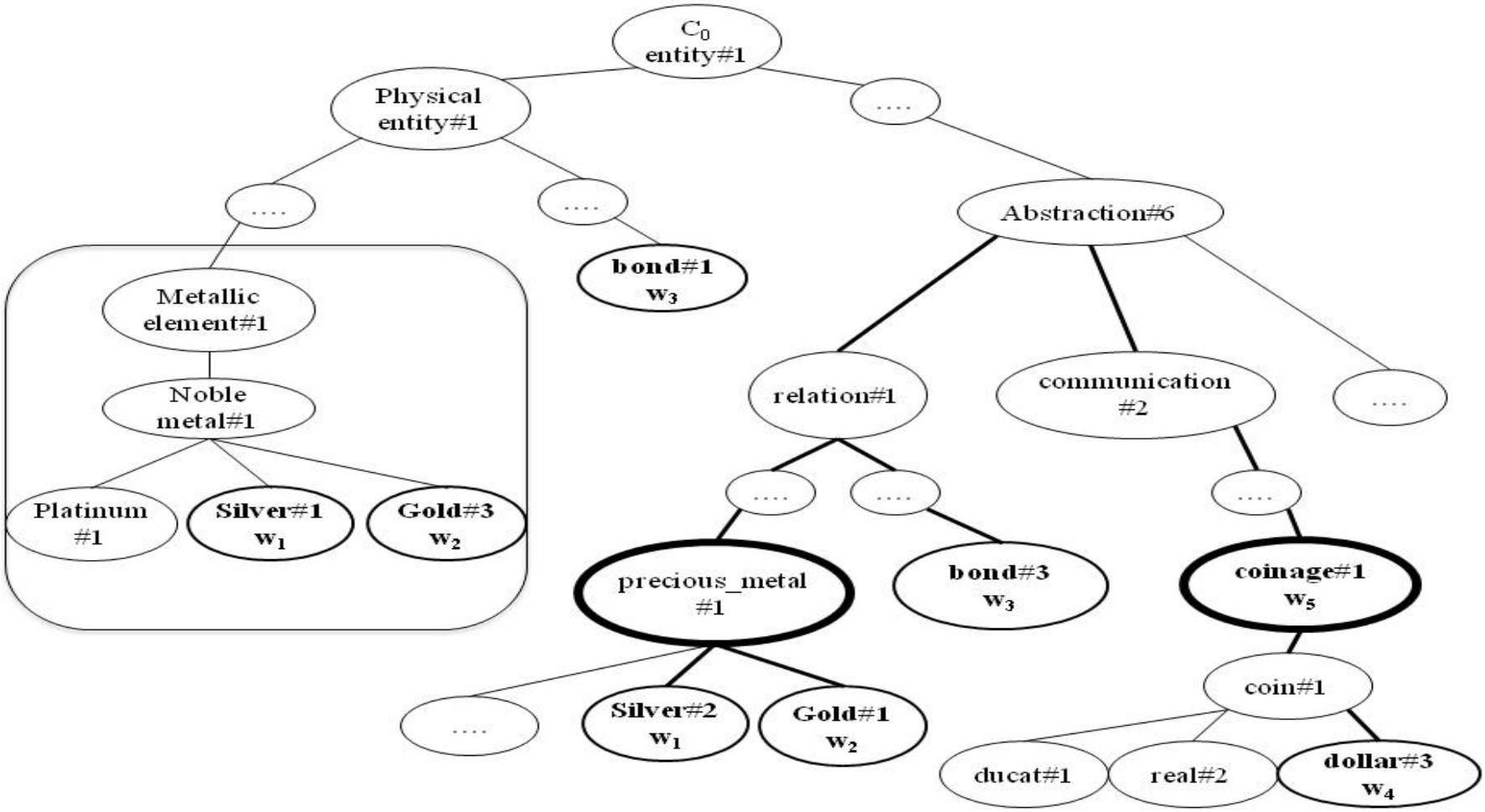
- Probabilistic method, each analysed document is a “mix” of topics, of decreasing probabilities.
- Three example topics :

```
père(1750.0) famille(1267.0) mère(1221.0) fils(1139.0)
  enfant(1088.0) jeune(771.0) grand(644.0)
  parent(589.0) ...

guerre(1074.0) armée(381.0) soldat(294.0)
  résistance(227.0) combat(214.0) général(211.0)
  officier(187.0) ...

hôpital(975.0) médecin(774.0) médical(549.0)
  service(407.0) santé(405.0) malade(363.0)
  docteur(306.0) ...
```

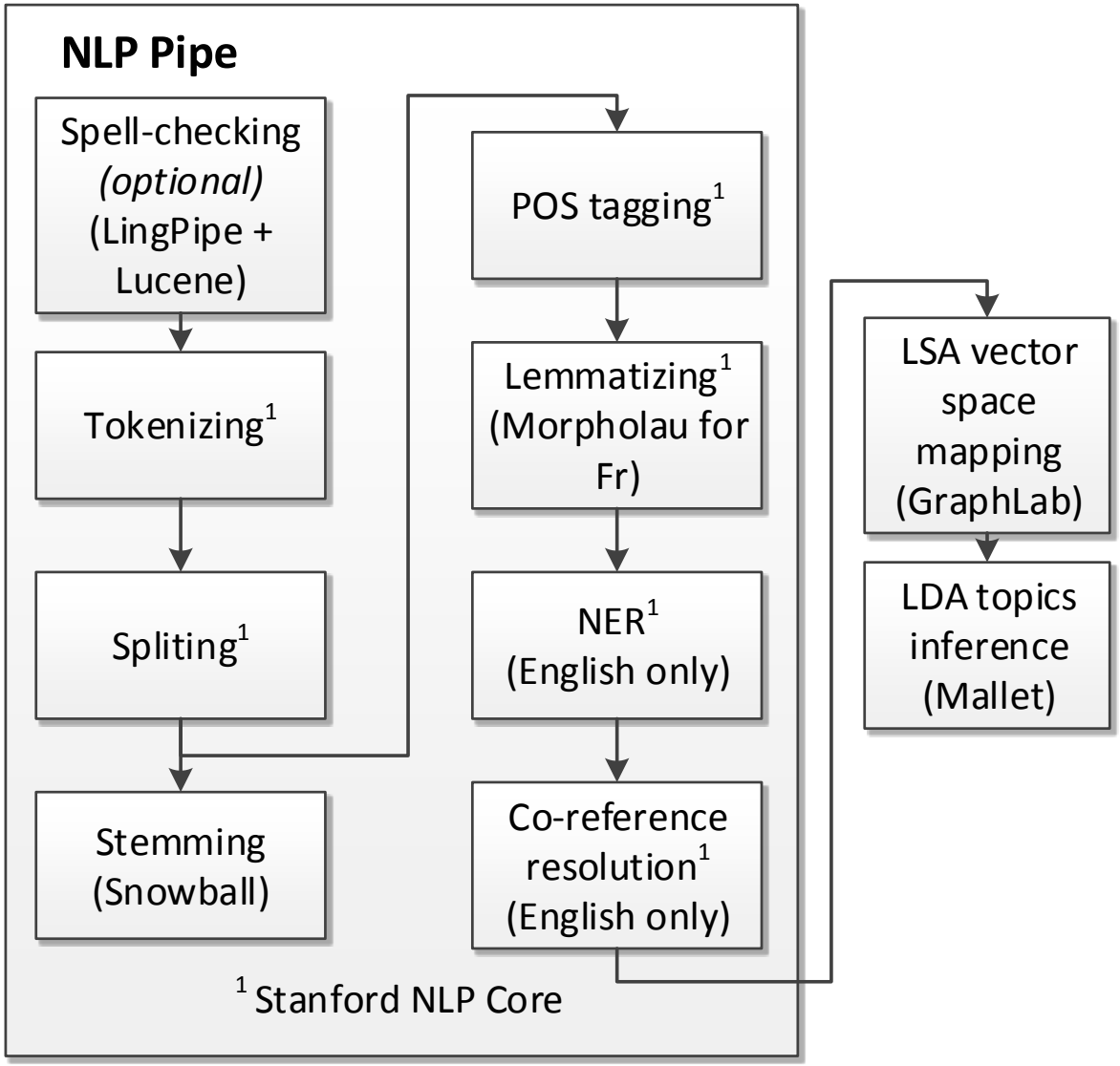
WordNet (wordnet.princeton.edu)

- Ontology > 150,000 concepts
- Nouns, verbs, adjectives, adverbs
- Built from psycholinguistics experiments → a semantic network of common concepts used in language

WordNet: Topical Subtrees – An Example



Text Pre-processing



Corpora Description

- 1,000 abstracts from Jensen & Grauwin database
- Abstracts from educational science domain
- Problems of scalability & processing of entire corpora using full NLP pipe tools

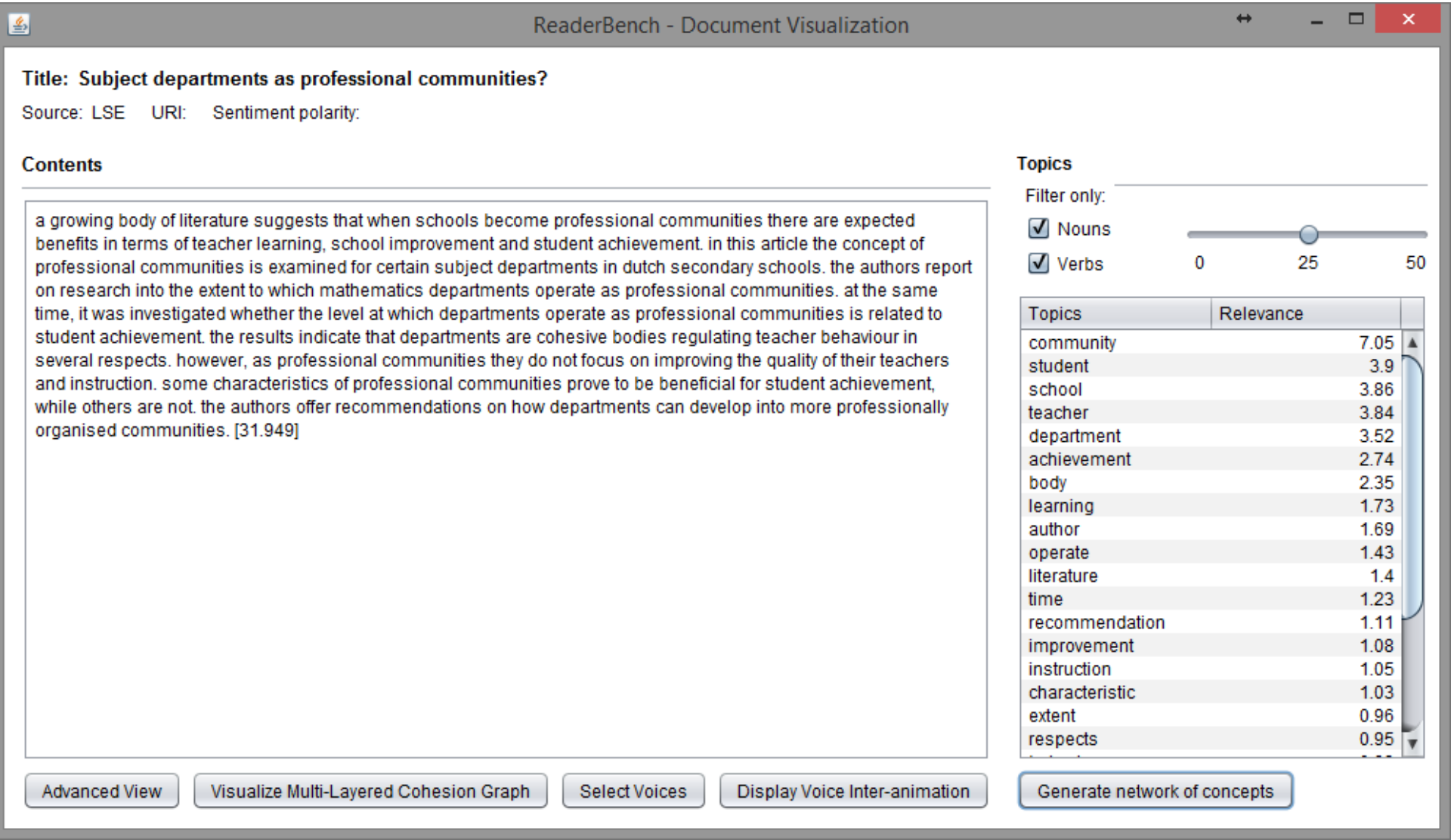
New Functionalities (1)

- Corpus similarity view
  - Conceptualize the corpora through semantic similarity between abstracts
  - Finds most similar pairs of articles
  - Determine most central articles
- Document centrality view
  - Find most similar articles to a selected article
  - Expand using a BFS strategy the graph by using the connected articles

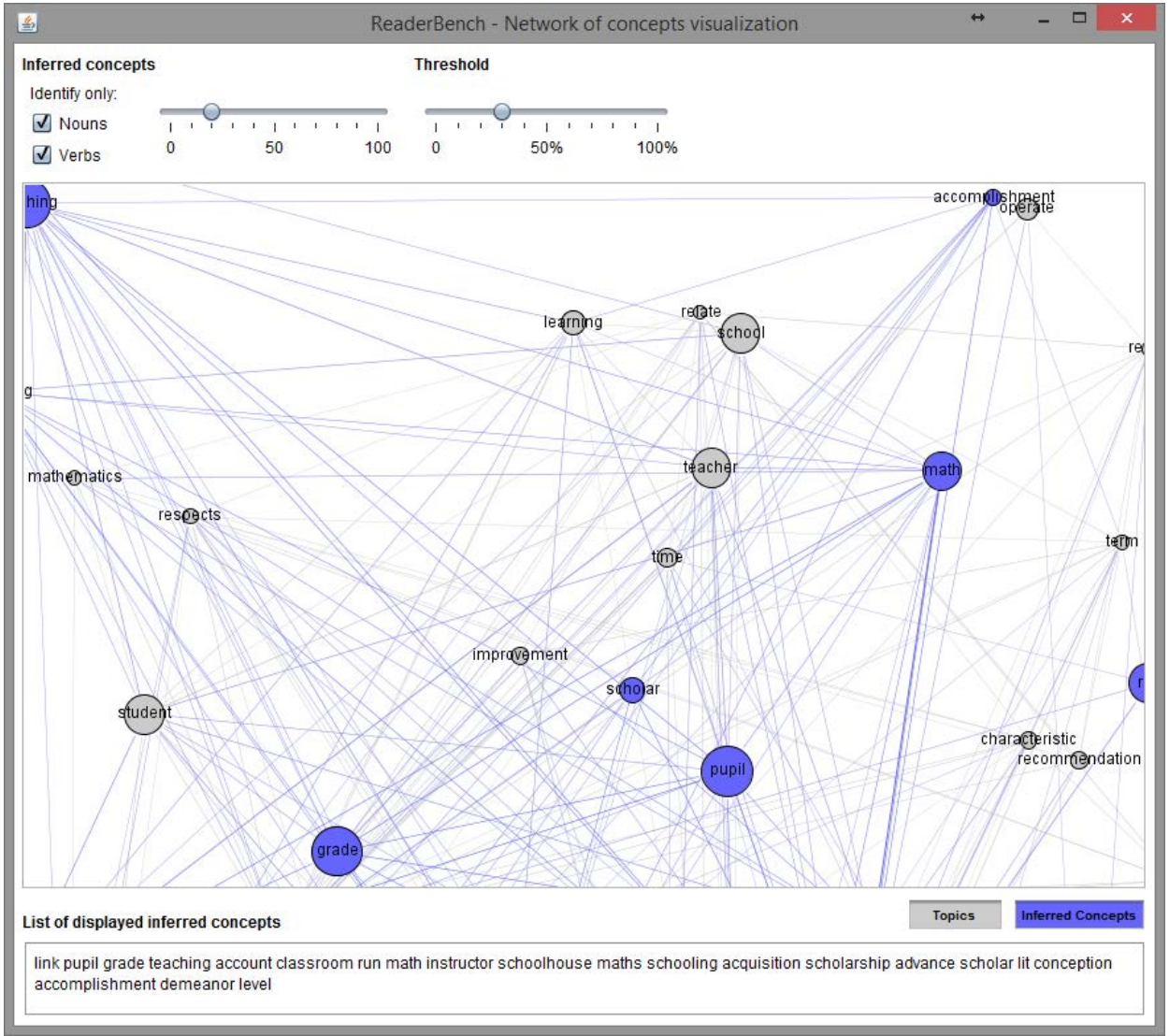
New Functionalities (2)

- Concept view
  - Identify most relevant concepts for the entire corpora
  - Relational meta models
- Keyword-abstract overlaps
  - Determine best overlap between keywords and abstract content
  - Compound score: occurrences of keywords (30%) and the semantic relatedness between keyword ”phrase” and abstracts (70%)

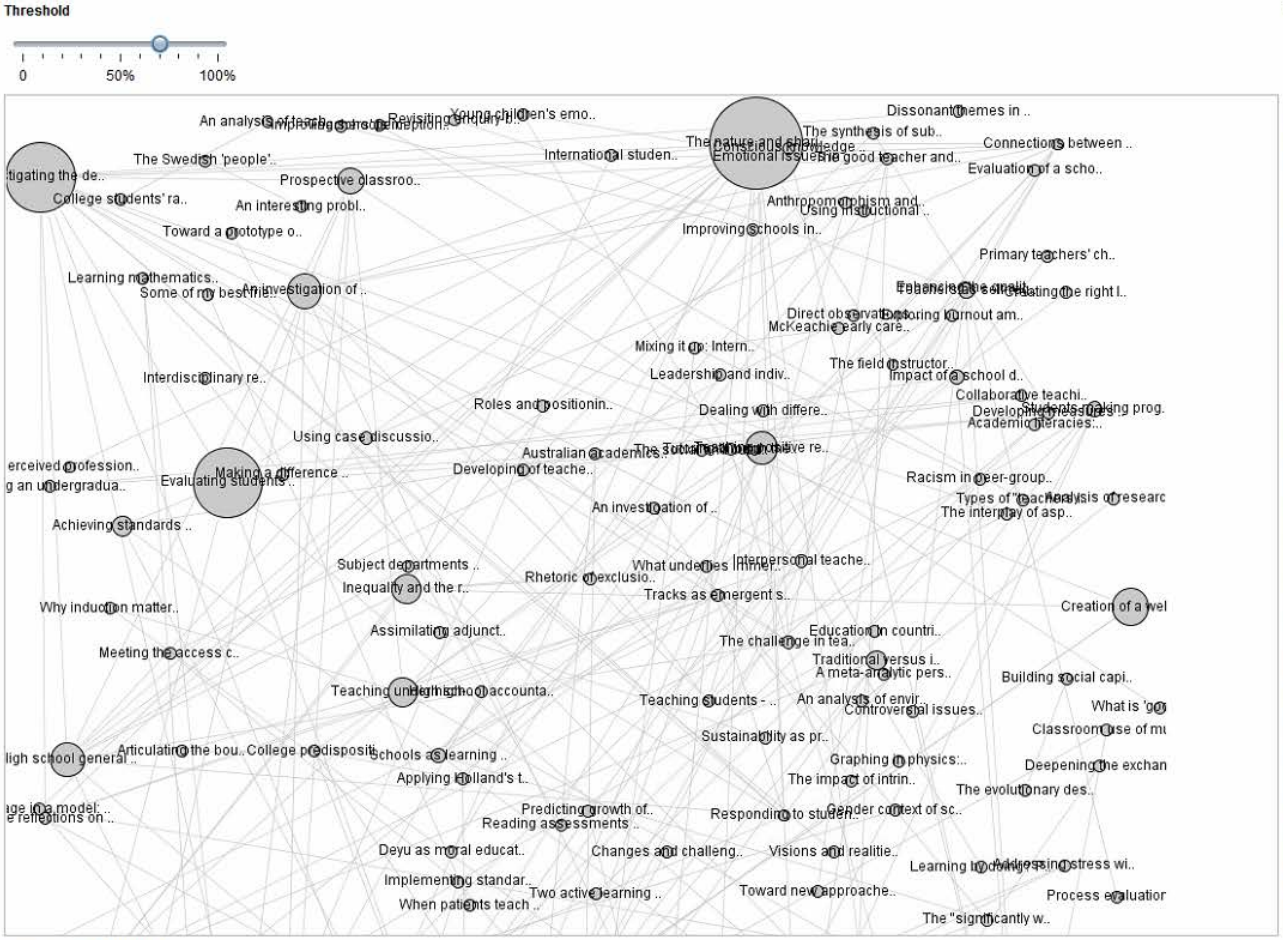
Article View



Concept View (LSA & LDA)



Corpus Similarities View



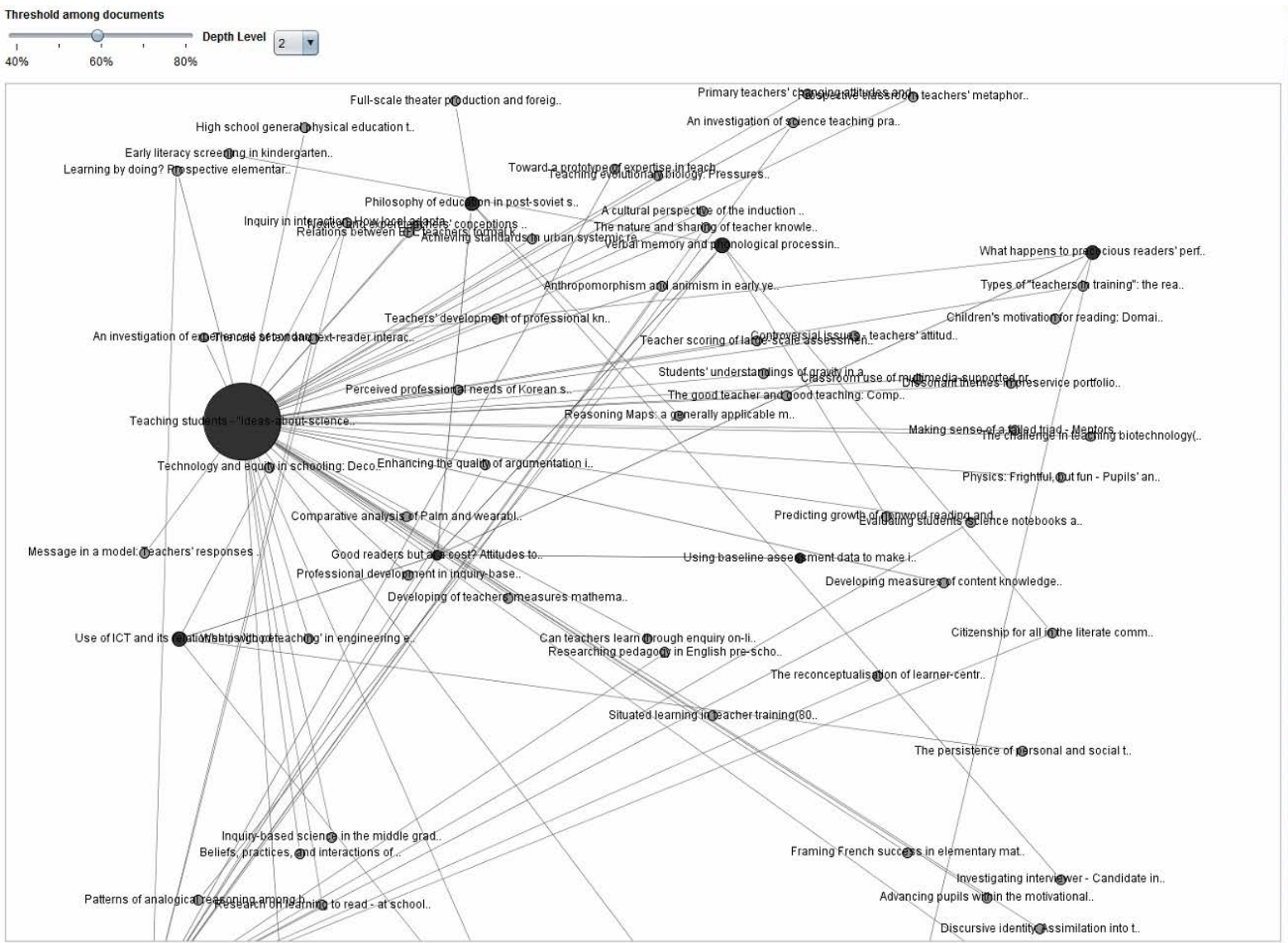
Distance to Central Document

Document	Distance
Novice and expert teachers' conceptions of learners' prior knowledge	0.77
Connections between teacher and student attitudes regarding script choic...	0.77
An investigation of science teaching practices in Indonesian rural seconda...	0.76
Message in a model: Teachers' responses to a court-ordered mandate for...	0.75
Investigating the decision-making skills of cooperating teachers and stud...	0.74
Using case discussion on the web to develop student teacher problem so...	0.73
South Korean high school English teachers' code switching: Questions an...	0.73
The good teacher and good teaching: Comparing beliefs of second-grade ...	0.72
Teaching evolutionary biology: Pressures, stress, and coping	0.72
The challenge in teaching biotechnology	0.72
High school general physical education teachers' behaviors and beliefs a...	0.71
Responding to students' views about university teaching: the experience of...	0.71
Relations between EFL teachers' formal knowledge of grammar and their ...	0.71
The field instructor as group worker: Managing trust and competition in gro...	0.71
The synthesis of subject and pedagogy for effective learning and teaching ...	0.70
Evaluating students' science notebooks as an assessment tool	0.70
Prospective classroom teachers' metaphorical images of selves and com...	0.70
Perceived professional needs of Korean science teachers majoring in che...	0.70
Some reflections on learning from early school experience	0.70
Analysing cognitive or non-cognitive factors involved in the process of phys...	0.70
Developing of teachers' measures mathematics knowledge for teaching	0.70
Developing measures of content knowledge for teaching reading	0.70
Interpersonal teacher behaviour and student outcomes	0.70
Primary teachers' changing attitudes and cognition during a two-year scie...	0.69
Impact of a school district's science reform effort on the achievement and ...	0.69
Inequality and the right to learn: Access to qualified teachers in California's...	0.69
Enhancing the quality of argumentation in school science	0.69
Controversial issues - teachers' attitudes and practices in the context of cit...	0.69
The interplay of aspirations, enjoyment, and work habits in academic ende...	0.68
Teacher scoring of large-scale assessment: professional development or...	0.68
Learning mathematics in a classroom community of inquiry	0.67
What is 'good teaching' in engineering education in India? A case study	0.67
Situated learning in teacher training	0.67
Visions and realities of Internet use in schools: Canadian perspectives	0.67
What underlies immersion students' production: The case of Avoir Besoin ...	0.67
Dissonant themes in preservice portfolio development	0.66
Achieving standards in urban systemic reform: An example of a sixth grad...	0.66
Tutoring through the internet: how students and teachers interact to constr...	0.66
Evaluation of a school-based intervention for HIV/AIDS prevention among ...	0.65
Between idealism and pragmatism: a case study of student teachers' ped...	0.65
Teaching under high-stakes testing - Dilemmas and decisions of a teach...	0.65
Toward new approaches for evaluating student field performance: Tappin...	0.65
Traditional versus integrated preservice teacher education curriculum - A c...	0.65
Teaching students - "Ideas-about-science": dimensions of effective practice	0.65

Top Similar Articles

Article 1	Article 2	Semantic Similarity Score
Developing of teachers' measures mathematics knowledge for teaching	Developing measures of content knowledge for teaching reading	0.86
Investigating the decision-making skills of cooperating teachers and student teachers of English in a Turkish context	Interpersonal teacher behaviour and student outcomes	0.79
South Korean high school English teachers' code switching: Questions and challenges in the drive for maximal use of English in tea...	Connections between teacher and student attitudes regarding script choice in first-year Japanese language classrooms	0.79
Gender context of schooling and levels of stress among early adolescent pupils	Gender organization of schooling and television viewing among early adolescents: a test of two alternative hypotheses	0.79
Teaching under high-stakes testing - Dilemmas and decisions of a teacher educator	Teachers as self-regulated learners	0.78
College students' ratings of student effort, student ability and teacher input as correlates of student performance on multiple-choice e...	An analysis of teachers' rating scales as sources of evidence for a standardised Greek reading test	0.77
The nature and sharing of teacher knowledge of technology in a student teacher/mentor teacher pair	Novice and expert teachers' conceptions of learners' prior knowledge	0.77

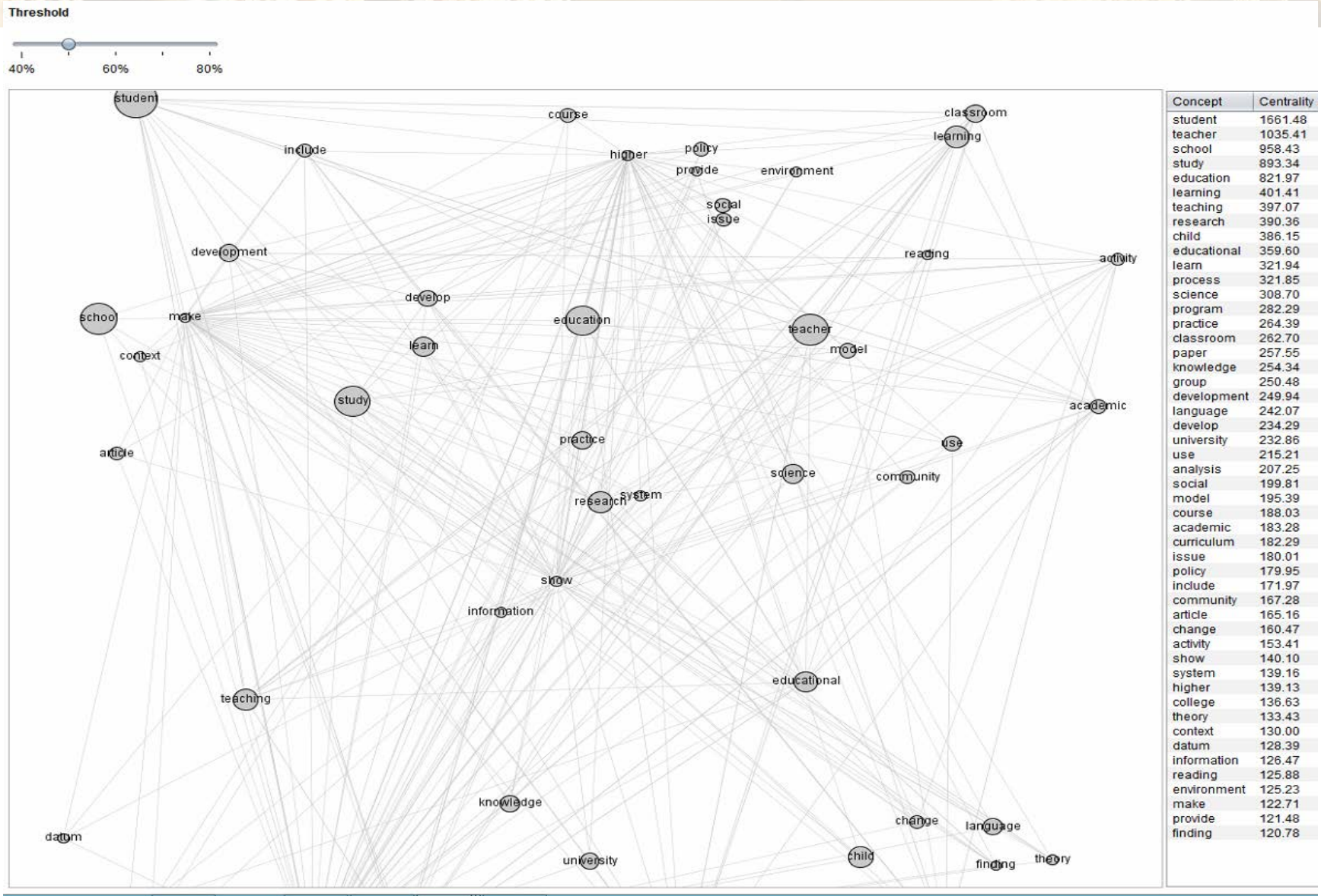
Document Centrality View



Top similar articles

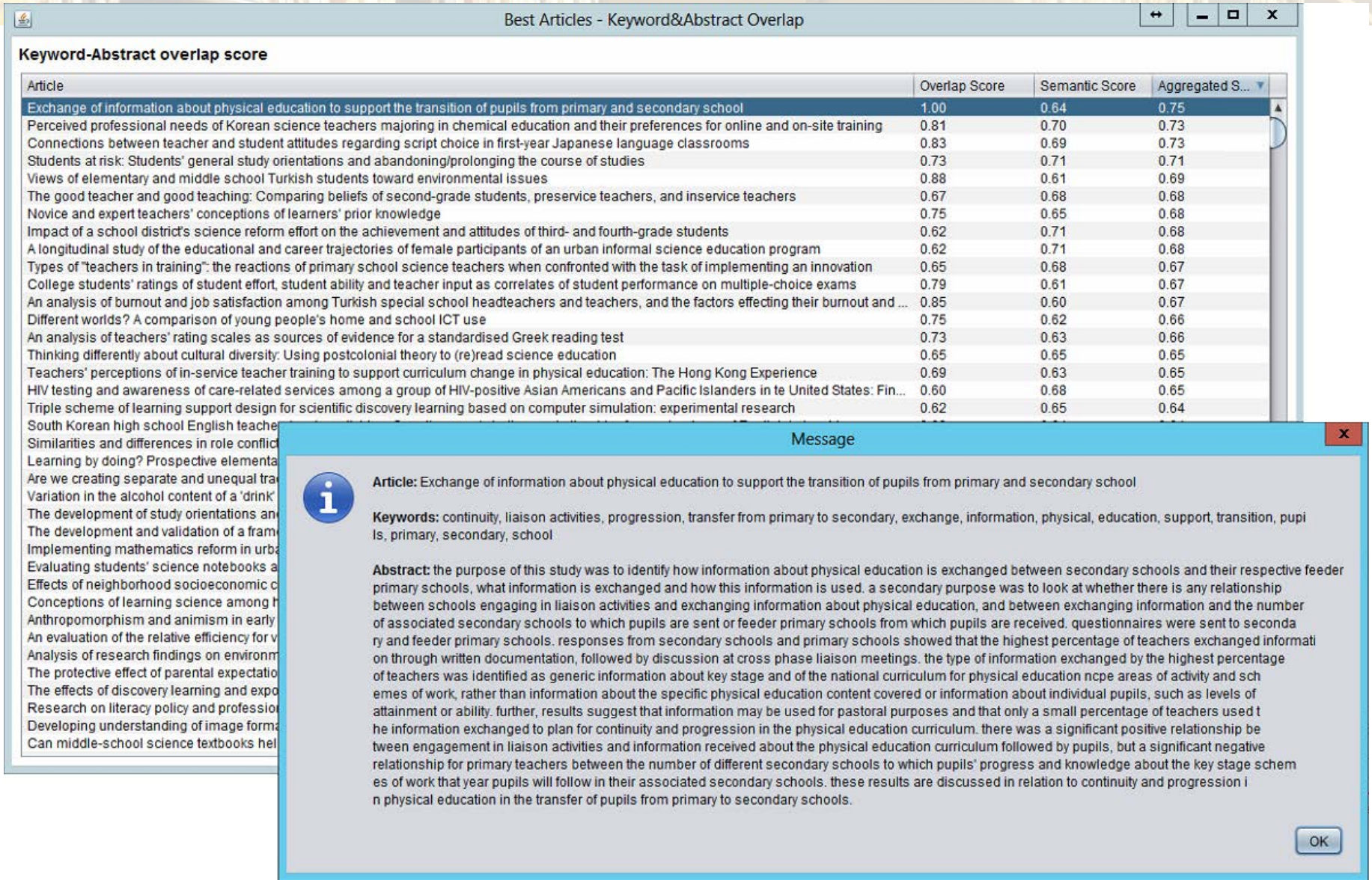
Article	Similarity
Verbal memory and phonological proce...	0.68
Philosophy of education in post-soviet s...	0.65
Use of ICT and its relationship with perf...	0.62
What happens to precocious readers' p...	0.61
Reading assessments in Kindergarten ...	0.61
Using baseline assessment data to ma...	0.60
Teaching students - "Ideas-about-scien...	0.59
Citizenship for all in the literate commun...	0.58
The persistence of personal and social ...	0.58
Patterns of analogical reasoning amon...	0.56
The role of text and text-reader interactio...	0.56
Early literacy screening in kindergarten: ...	0.56
Predicting growth of nonword reading a...	0.55
Framing French success in elementary ...	0.55
Developing measures of content knowl...	0.55
Research on learning to read - at schoo...	0.55
Researching pedagogy in English pre-s...	0.54
Situated learning in teacher training	0.52
Students' understandings of gravity in a...	0.51
Beyond phonological skills: broader lan...	0.51
Children's motivation for reading: Doma...	0.51
Investigating interviewer - Candidate int...	0.50
Anthropomorphism and animism in earl...	0.50
The challenge in teaching biotechnology	0.50
Can teachers learn through enquiry on-l...	0.50
Impact of a school district's science refo...	0.50
Inquiry in interaction: How local adaptati...	0.50
Teachers' development of professional ...	0.49
Enhancing the quality of argumentation i...	0.49
Relations between EFL teachers' formal...	0.49
A study of science teaching self-efficacy ...	0.49
Professional development in inquiry-ba...	0.48
Primary teachers' changing attitudes an...	0.48
Evaluation of a school-based interventio...	0.48
Prospective classroom teachers' metap...	0.48
Message in a model: Teachers' respon...	0.47
An investigation of science teaching pra...	0.47
Developing of teachers' measures math...	0.47
Beliefs, practices, and interactions of te...	0.47
Evaluating students' science notebooks...	0.47
Full-scale theater production and foreig...	0.47
Inquiry-based science in the middle gra...	0.47
Novice and expert teachers' conception...	0.47
Classroom use of multimedia-supporte...	0.46
Teacher scoring of large-scale assess...	0.46
Toward a prototype of expertise in teachi...	0.46
Learning by doing? Prospective elemen...	0.46
Achieving standards in urban systemic r...	0.45
Types of "teachers in training": the reacti...	0.45
The difficult relationship between theory ...	0.45
Physics: Frightful, but fun - Pupils' and t...	0.44
The nature and sharing of teacher know...	0.44
High school general physical education ...	0.44
Technology and equity in schooling: De...	0.44
Comparative analysis of Palm and wear...	0.44
The good teacher and good teaching: C...	0.44
Teaching evolutionary biology: Pressure...	0.43

Concept View



Concept	Centrality
student	1661.48
teacher	1035.41
school	958.43
study	893.34
education	821.97
learning	401.41
teaching	397.07
research	390.36
child	386.15
educational	359.60
learn	321.94
process	321.85
science	308.70
program	282.29
practice	264.39
classroom	262.70
paper	257.55
knowledge	254.34
group	250.48
development	249.94
language	242.07
develop	234.29
university	232.86
use	215.21
analysis	207.25
social	199.81
model	195.39
course	188.03
academic	183.28
curriculum	182.29
issue	180.01
policy	179.95
include	171.97
community	167.28
article	165.16
change	160.47
activity	153.41
show	140.10
system	139.16
higher	139.13
college	136.63
theory	133.43
context	130.00
datum	128.39
information	126.47
reading	125.88
environment	125.23
make	122.71
provide	121.48
finding	120.78

Keywords-abstract Overlap



3. FURTHER WORK

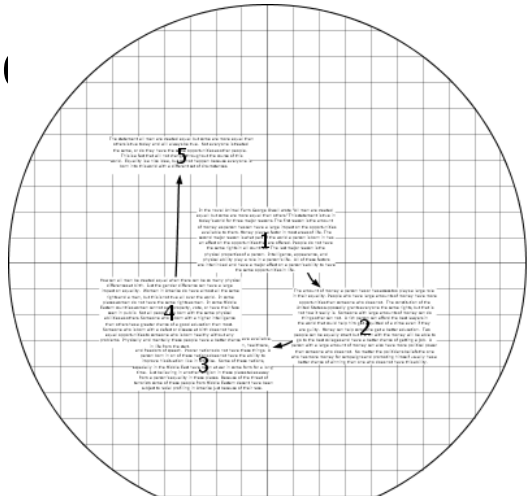
Towards a *Content*-based Citation Analysis (CAA) [Ding et al. 14]

There is at least 15 reasons why to refer to a paper (practices, purpose, or motivation) [Garfield 64]. Full-text analyses allow to dig deeper into these motivations

- Syntactic CAA: Citation Proximity Analysis [Gipp & Beel 09]: accounts for co-occurrences within sentences, paragraphs or paper sections (contexts) [Elkiss et al. 08]
- Semantic CAA: Account for semantic similarity across full texts

Go Full Text!

- Semantic analyses of full text:
  - Use sentiment analysis to dectect authors' citation motivation
  - Inter-paragraph cohesion to display overall paper's semantic flow [O'Rourke & Calvo 09]
- How to decrease the computational demand?
  - Consider using either LSA *or* LDA instea



Warning by Icomatic http://thenounproject.com/term/warning/

Further Work—Collaboration

- Domain representation readability (semiotic maps)
- Server setup to play with data
- Focus on specific sub-fields, like TEL -> ANR Orphee
- Merge our 2 approaches
  - 1st tier: Semantically compare the summaries-> Map
  - 2nd tier: Adjust the map with regard to papers' bib coupling

Further Work—Likely Milestones

- Dynamic impact assessment: effect of the introduction of bulks of published papers on the paper's space (over years)
- LDA-based measurement [Song & Ding 14]
  - to tag clusters
  - to extract key-words and compare with the actuals
- Use researchers' networks and classification (Mendeley, CiteULike, etc.), a.k.a. alt-metrics, to uncover sub-disciplines [Kraker et al., in press]

Even Further

- Recommender systems (which paper to read)
- Writing advices (how to write: coherence, which paper to cite)
- Citation genetic genealogy: a cited paper is the "son" of the papers citing it and shares their "ADN" [Sun & Zhu 12]
- Adding RDF papers metadata (e.g., CiTO tags http://purl.org/spar/cito/ & SPAR, Semantic Publishing and Referencing initiative)

Discussion

- Bib. coupling and semantic analysis are complementary approaches
- Relate computer-based analyses with human-based ones (results acceptability)

THANK YOU FOR YOUR ATTENTION!

Refs available at

http://www.citeulike.org/user/pdessus/tag/educmap

Contact

trausan@gmail.com

philippe.dessus@upmf-grenoble.fr

paraschiv.ionut@gmail.com

mikedascalu@yahoo.com

Aggregated Cohesion Function

- Normalized value in [0; 1] considering the following:
  - Inverse normalized distance
  - Lexical proximity – identical lemmas and semantic distances in WordNet / WOLF (Path, *Wu-Palmer*, Leacock-Chodorow)
  - Semantic similarity
    - Cosine similarity – LSA (GraphLab)
    - Jensen – Shannon Divergence – LDA (Mallet)

Topics Extraction

- Individual normalized term frequency - *Statistical presence*
- Semantic similarities (LSA + LDA) - *Semantic relatedness*
  - Analysis element (sentence > block > document)
  - Entire document
- Weighted similarity with semantic chain - *Semantic coverage*

Topic Relevance