SEMANTIC INDEXING
FOR INFORMATION RETRIEVAL AND BIBLIOMETRIC ANALYSIS

Rob Koopman
Shenghui Wang
OCLC

2016 ASIS&T Annual Meeting
17/10/2016

A common ground

- Information retrieval (IR) is the activity of obtaining information resources **relevant** to an information need from a collection of information resources
- Bibliometrics is statistical analysis of written publications

Relatedness or similarity is the base

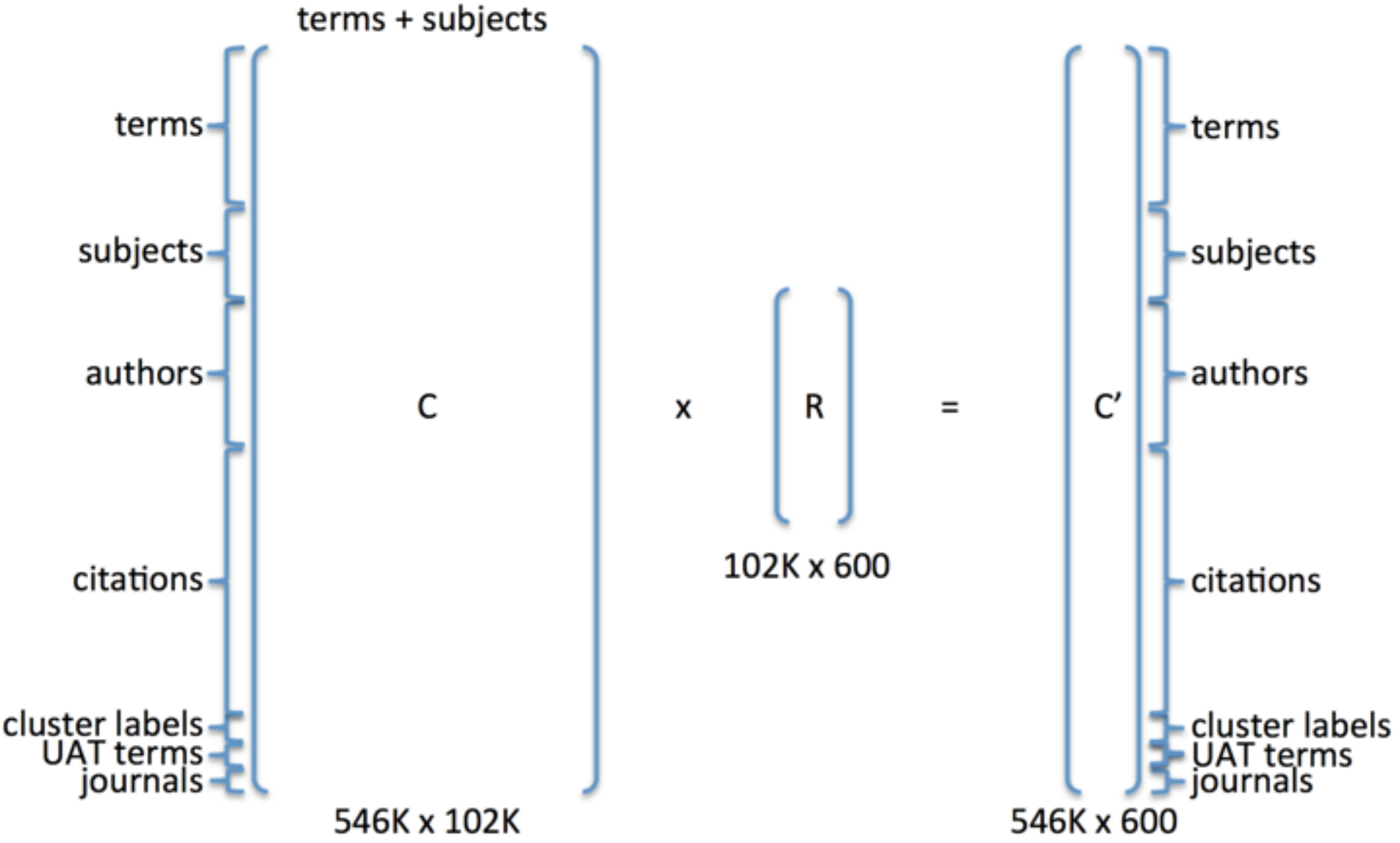
Build semantic representation

- *Statistical Semantics* based on the assumption of "a word is characterized by the company it keeps" [firth1957]
- *Distributional Hypothesis:* words that occur in similar contexts tend to have similar meanings. [harris1954, sahlgren2008]

Word embedding

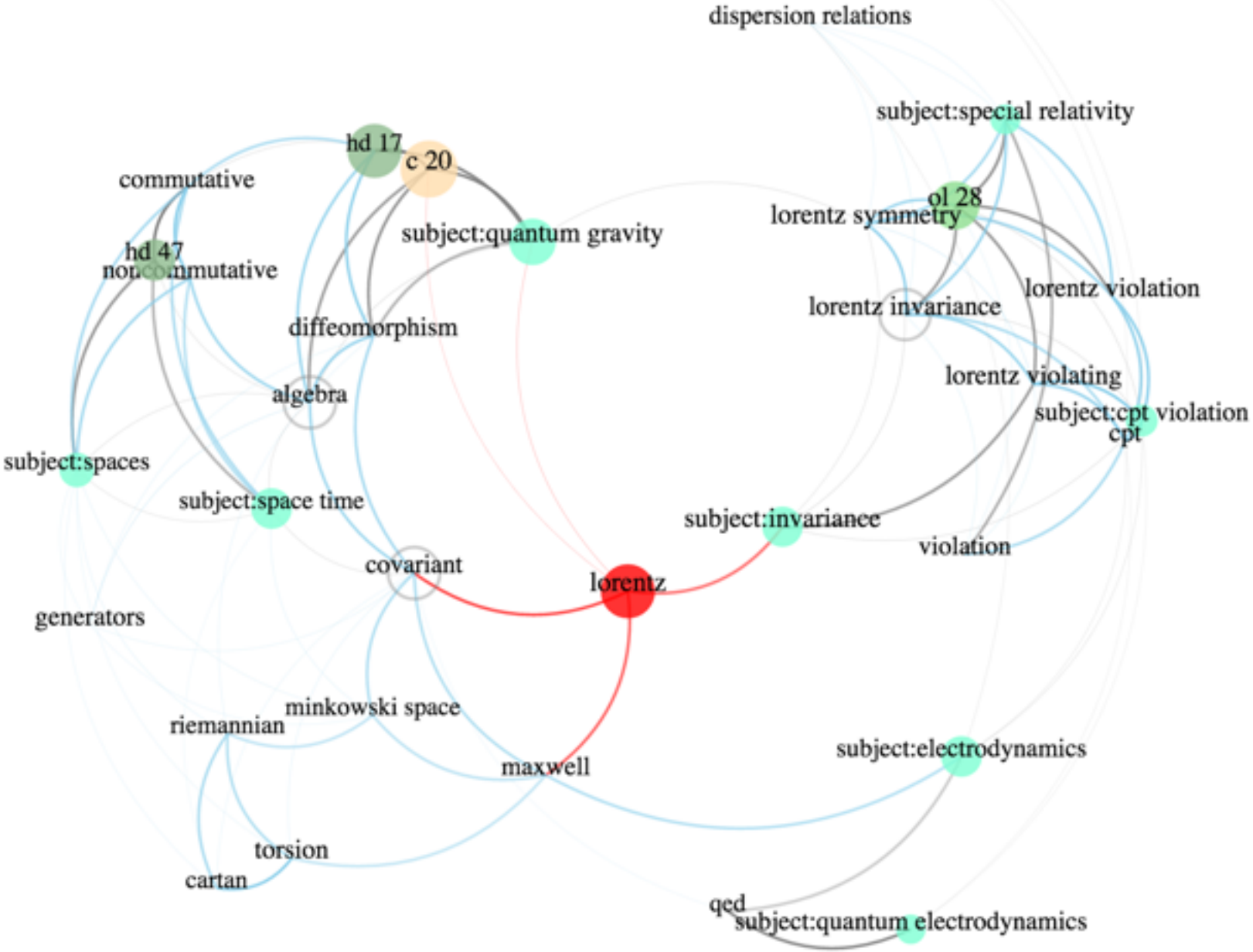
- Words or phrases => vectors of real numbers
- Methods:
  - Explicit representation of the context, e.g. word-document/word-word co-occurrence matrix
  - Dimension reduction (LSA and its variations and extensions, Random projection, etc.)
  - Neural networks (word2vec)

Random projection



From semantic representation

- Similarity calculation (cosine)
- Context visualisation



http://thoth.pica.nl/astro/relate?input=lorentz

Different techniques lead to different embeddings

Word	Method	Top 10 similar words/phrases (trained on 25 million Medline articles)
Frog	Random projection	rana, isolated frog, frog sartorius muscle, frogs, frog rana, frog muscle, liagushki, amphibian, liagushek, frog sartorius muscles
	word2vec	toad, bullfrog, amphibian, rana, frogs, turtle, salamander, caudiverbera, bufo, leptodactylid
	doc2vec	toad, crayfish, bullfrog, amphibian, rabbit, chicken, rat, ferret, frogs, goldfish
Brain	Random projection	brain areas, brain regions, brains, cortex hippocampus, rodent brain, thalamus hippocampus, cerebrum, mamalian brain, cortex thalamus, cortex
	word2vec	brains, cerebral, cns, cerebellum, cerebrum, brainstem, hippocampus, forebrain, neocortical, cortical
	doc2vec	liver, kidney, cns, skin, testis, brainstem, lung, peripheral, myocardium, brains

Computational cost

Method	Time	Thread
Random projection	2:42	1
Word2Vec	10:38:34	16
Doc2Vec	22:27:45	16

From words to documents

What about searching for articles most related to this statement:

radiotherapy is associated with significantly more complications in the presence of a breast reconstruction

From words to documents

What about searching for articles most related to this statement:

radiotherapy is associated with significantly more complications in the presence of a breast reconstruction

- Bag of words or TF-IDF
- LSA, LDA
- Neural language model
- Word mover's distance
- **Weighted average**

Application in topic delineation

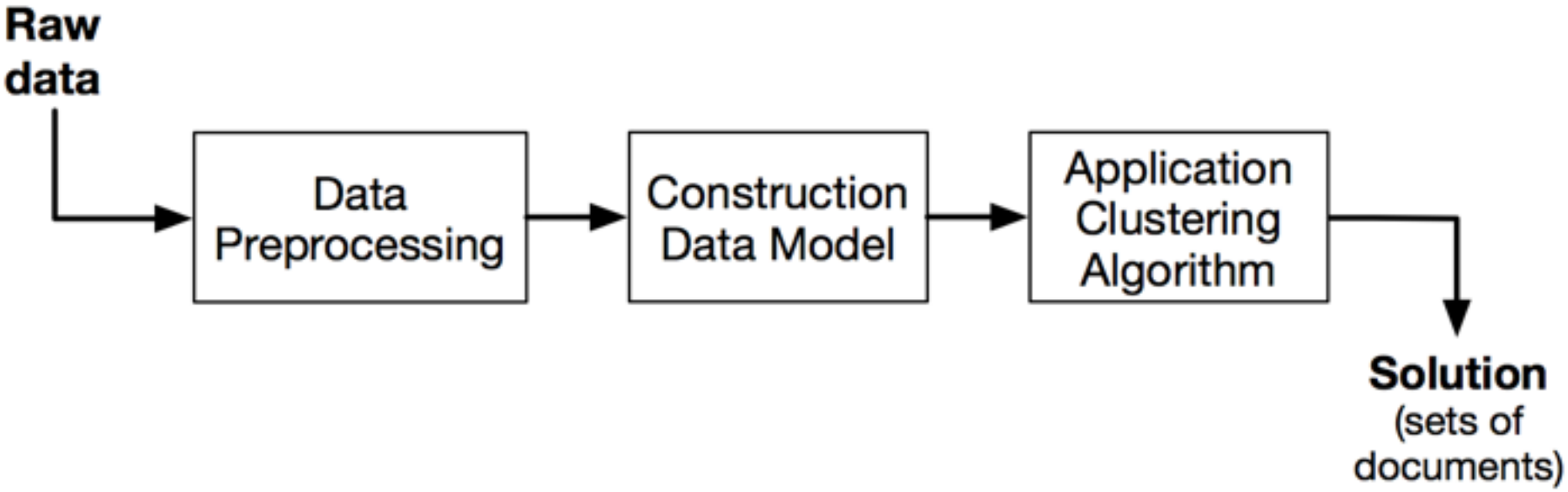


Table 1 Combinations of data models and clustering algorithms

	direct citation	bibliographic coupling (bc)	hybrid (bc + NLP terms)	semantic matrix	global direct citation map
infomap	**u**	–	–	–	–
SLMA	**c**	–	–	–	**sr**
Memetic	**sr**	–	–	–	–
Louvain	–	**eb**	**en**	**ol**	–
k-means	–	–	–	**ok**	–

Semantic indexing is comparable

Table 3 Normalised Mutual Information (emphasis: **max**, *min* value)

	sr	c	u	ok	ol	en	eb
sr	1.00	0.36	0.37	0.33	0.33	*0.24*	0.31
c	0.36	1.00	0.63	0.46	0.52	0.32	0.38
u	0.37	**0.63**	1.00	0.42	0.47	0.30	0.36
ok	0.33	0.46	0.42	1.00	0.52	0.33	0.36
ol	0.33	0.52	0.47	0.52	1.00	0.31	0.36
en	0.24	0.32	0.30	0.33	0.31	1.00	0.33
eb	0.31	0.38	0.36	0.36	0.36	0.33	1.00

Application in updating medical guidelines

- Medical guidelines need to be updated with the recent publications
- With reasonable word embedding and smart weighing techniques, we can exclude 99.9% articles in Medline
- However, it is not enough for medical doctors, as they only have time to look at the top 30 hits
- Maybe bibliometric analysis can help here, e.g. citations, impact factors, authors, etc

Messages to take home

- Semantic indexing is old and new
- Many applications
- Best in combination with other methods

People are working on it too

- Aggregating Continuous Word Embeddings for Information Retrieval, CVSC@ACL '13
- Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data, CIKM '13
- Learning to Reweight Terms with Distributed Representations, SIGIR '15
- Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings, SIGIR '15
- Word Embedding Based Generalized Language Model for Information Retrieval, SIGIR '15
- Integrating and Evaluating Neural Word Embeddings in Information Retrieval, ADCS '15
- From Word Embeddings to Document Distances, ICML '15
- Short Text Similarity with Word Embeddings, CIKM '15
- A Dual Embedding Space Model for Document Ranking, WWW '16
- Improving Language Estimation with the Paragraph Vector Model for Ad-hoc Retrieval, SIGIR '16
- Embedding-based Query Language Models, ICTIR '16
- Estimating Embedding Vectors for Queries, ICTIR '16