What we learned about

LSI for Information Retrieval

and

Automatic Fill-in of Missing Metadata

Ying Wang

My focus this summer was on …

1. Building a generic search interface
   - Based on Latent Semantic Indexing (LSI) - *to be explained!*
   - With state-of-the-art algorithmic components (BM25)

2. Exploring approaches for filling in missing metadata

And apart from this, also analysis of archives as David reported on

1. Searching: what is this LSI?

A standard algorithm for information retrieval

- Developed in 1990 by Susan Dumais at Bell Labs
- Can give “magical” results.
- Can give terrible results.

The power of LSI is that it

- can find hidden semantic structure
- increase recall

An illustrative example *Berry and Dumais, 1995*

related through "algorithm"

B1	A course on integral equations
B2	Attractors for semigroups and evolution equations
B3	Algorithms: theory, implementation and applications
B4	Aspects of partial differential equations
…..	…….
B7	Knapsack problems: Algorithms and computer implementations
…..	…….
B11	Oscillation theory for neutral differential equations with delay
…..	…….
B17	MB type integrals and their applications to convolution theory

Find documents that best match "application and theory"

Exact matches only return B3, B11, B12 and B17

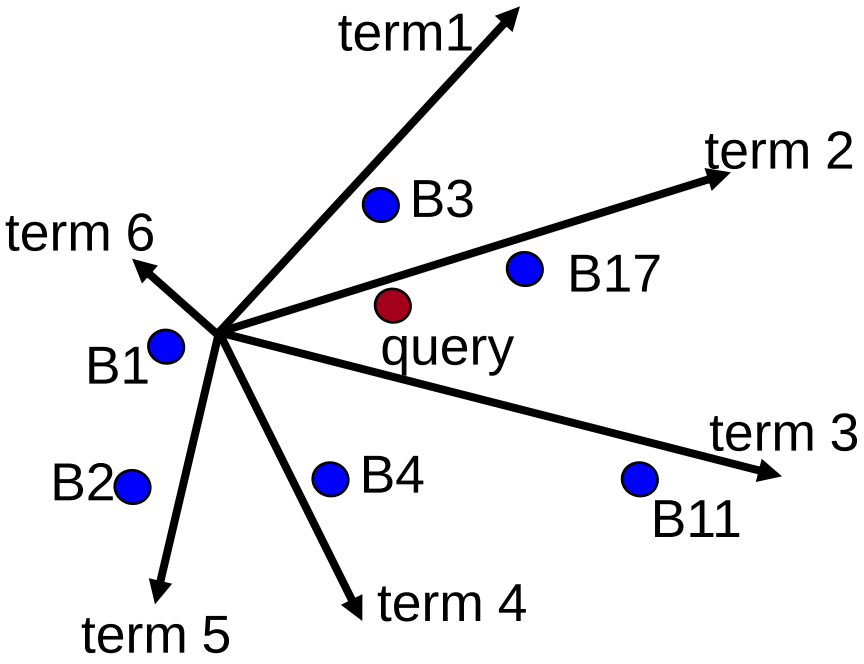
LSI starts by forming a matrix

	B1	B2	B3	B4	…	B7	B8	…	B11	…	B17
algorithms	0	0	1	0	…	1	0	…	0	…	0
application	0	0	1	0	…	0	0	…	0	…	1
delay	0	0	0	0	…	0	0	…	1	…	0
differential	0	0	0	1	…	0	1	…	1	…	0
:	:	:	:	:		:	:		:		:
systems	0	0	0	0	…	0	1	…	0	…	0
theory	0	0	1	0	…	0	0	…	1	…	1

The *term-document matrix A* stores weights a_{ij} for term i in document j

Find columns (docs) that best match query vector $q = [0 \quad 1 \quad 0 \quad \ldots \quad 0 \quad 1]^T$

Super high-dimensional search

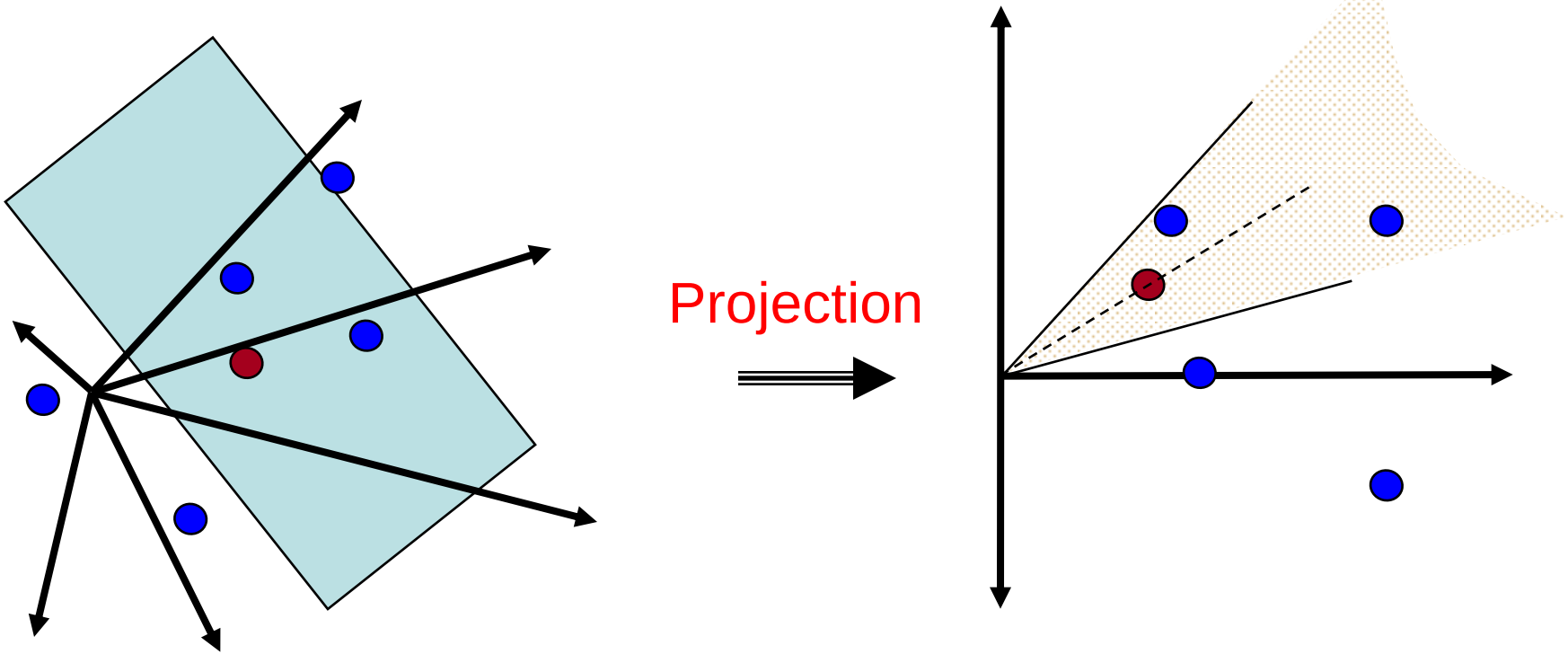


Find document vectors "close" to query vector in m-dimensional space

(with m the number of possible terms)

Measure of closeness is given by angle between vectors

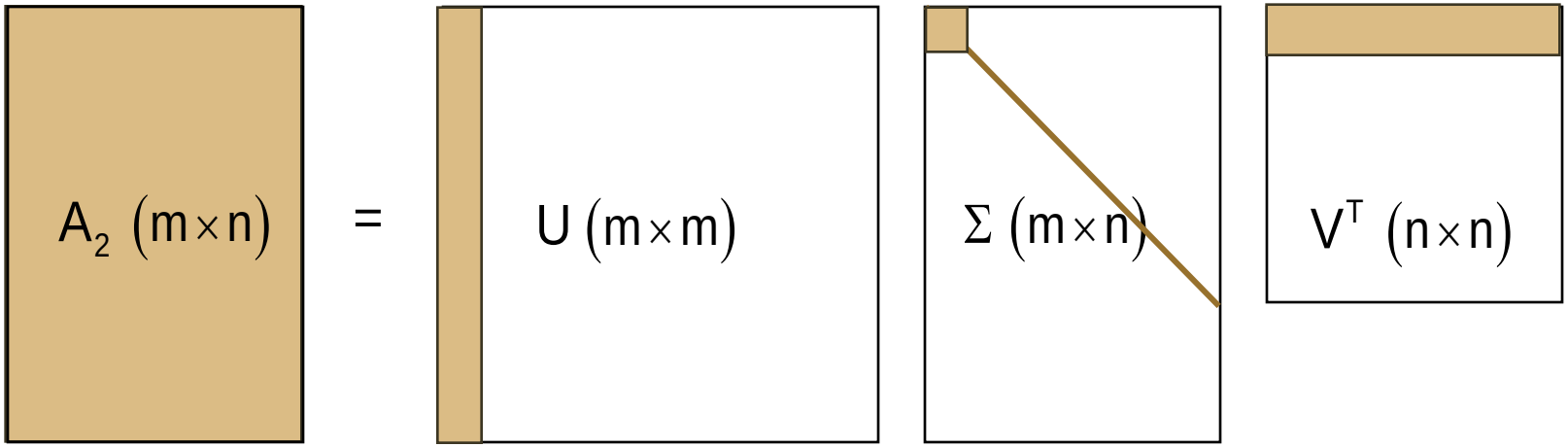
Project on lower-dimensional space



How should we construct this projection???????

SVD to the rescue! *Deerwester, Dumais, et al. (1990)*

Application of the singular value decomposition (SVD)

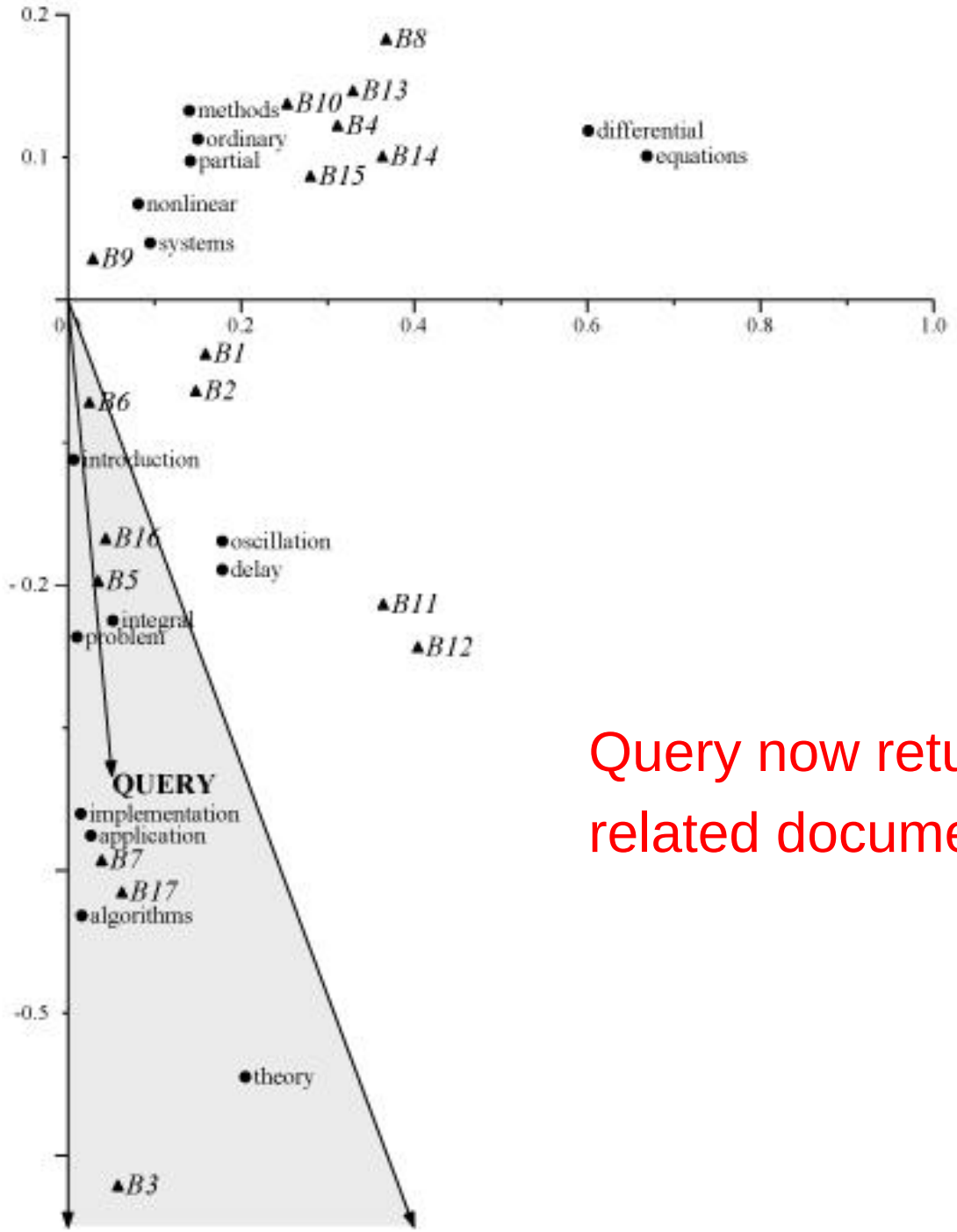


$$A = U\Sigma V^T = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + \sigma_3 u_3 v_3^T + \cdots + \sigma_p u_p v_p^T$$

σ decreasing

Work in the plane spanned by the remaining vectors u_i

Locations of terms and documents in plane easy to find



Query now returns related documents also

Interesting research questions

Is LSI effective on small (<10 words) documents?

e.g., LCSH, OAI titles.

Can we use LSI for multilingual search?

How should it be adapted / extended? (see Vinayak's presentation)

LSI on Small Documents

DEMO!

LSI on Small Documents

~~DEMO!~~

It's that bad.

LSI on Small Documents

What did we learn?

For LCSH and OAI title, matrix decomposition algorithms are not likely to be helpful.

2. Fill-in of missing metadata

Not all metadata are created equally

- Fields may be missing, incomplete, or erroneous

How many bibliographic records are incomplete?

- American Memory, Non MARC ≈ 10%, e.g. about 1 year at 10 min/entry

Why this matters?

- reduce search and browse quality
- quality of visualization of results (history line, geographical displays)

What we want to build

Tool to automatically and accurately fill-in missing metadata

- Automation saves labor, and is (potentially) fast and scalable

Automated fill-in is an interesting research topic

- Open area of research
- many potential applications
- error-correction tool

Missing Data Here?

<indexing_data_id>**cic**</indexing_data_id>

<item_title>**no. 31 [cover]: Our Newest Citizens: Chinese native sons registering at the City Hall: From The Wave: v. 15, Jan. - Dec. 1896**</item_title>

<author_creator>**Wave Publishing Company**</author_creator>

<source_collection>**The Wave: v. 15, Jan. - Dec. 1896**</source_collection>

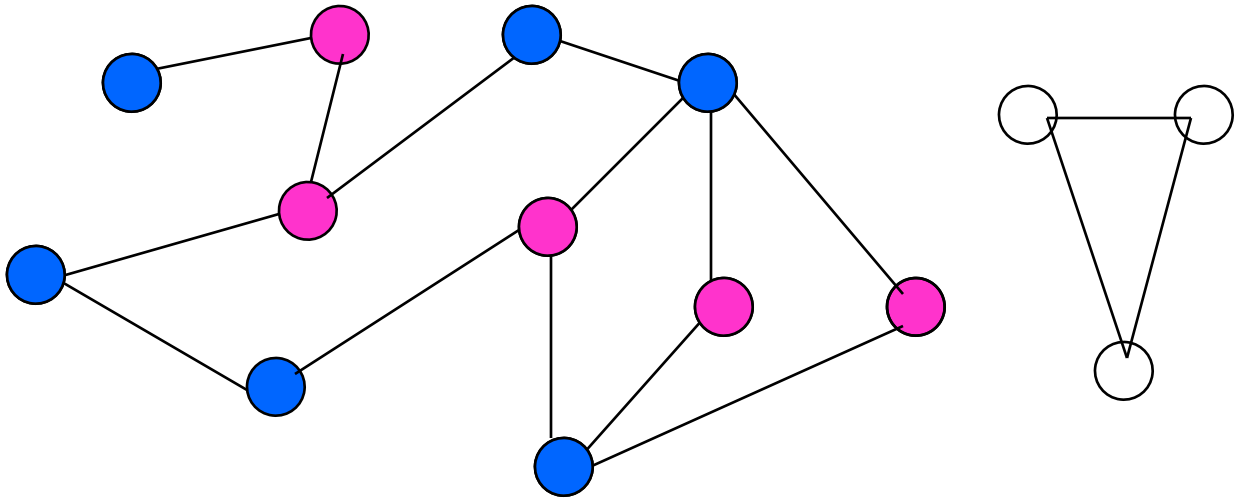
<collection_id>**cubcic**</collection_id>

<document_type>**still image**</document_type>

<genre authority="bgtchm">**Photographs**</genre>

<text_date>**San Francisco**</text_date>

Boys and girls



Each node is a child. Boys are blue and girls are pink.
Each pair of friends is linked by an edge.
Each child can have at most one friend from his or her gender.

What do we learn from the example?

We have:

An incomplete set of data.

A set of rules on the data.

Algorithm:

Repeatedly apply the rules to deduce new data.

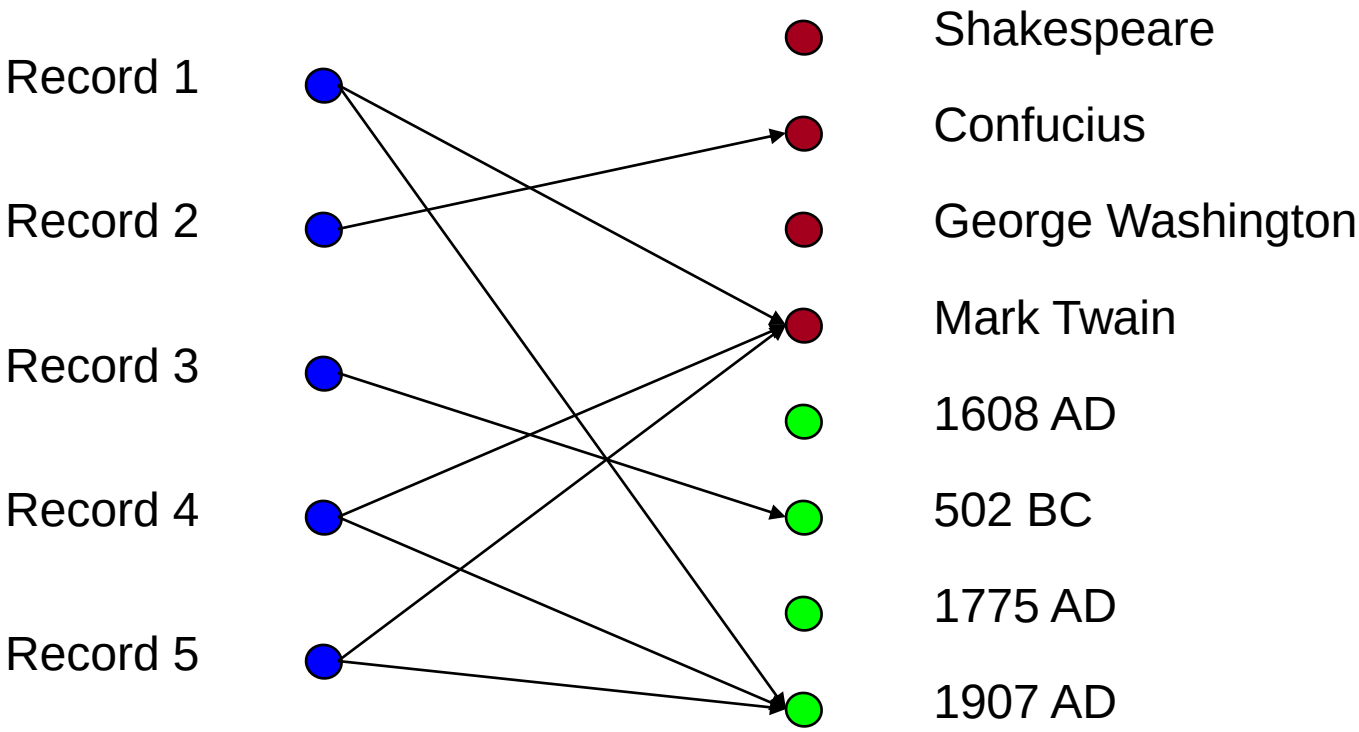
What are the rules for metadata?

Interaction between different fields.

Explicit interaction: Like author and date. Two works of the same author must be close in dates.

Implicit interaction: Also like author and date. Usually all works of an author are created in a shorter period. Such period can be found by statistical analysis.

A bipartite graph model



Records are vertices on the left. Possible outcomes of metadata fields are on the right. Information propagates from nodes to their neighbors.

Algorithm

- Retrieve information from related record
- Manually define some rules about the interaction of different data fields
- Learn hidden rules by machine learning algorithms
- Repeatedly apply above operations until no more information can be deduced

Conclusions and Future Directions

- LSI does not work directly on LCSH or OAI.
  - Look at multilingual LSI for WDL
  - Improve LSI to work on short documents
  - Use other techniques like BM25

- Missing metadata problem is a promising research problem that needs further investigation.
  - Collect all information
  - Try existing methods
  - Try new methods