Detecting Topic Evolution in Bibliographic Database Exploiting Citations

Graduate School of
Systems and Information Engineering,
University of Tsukuba

Hirotoshi Ito
Toshiyuki Amagasa
Hiroyuki Kitagawa

Outline

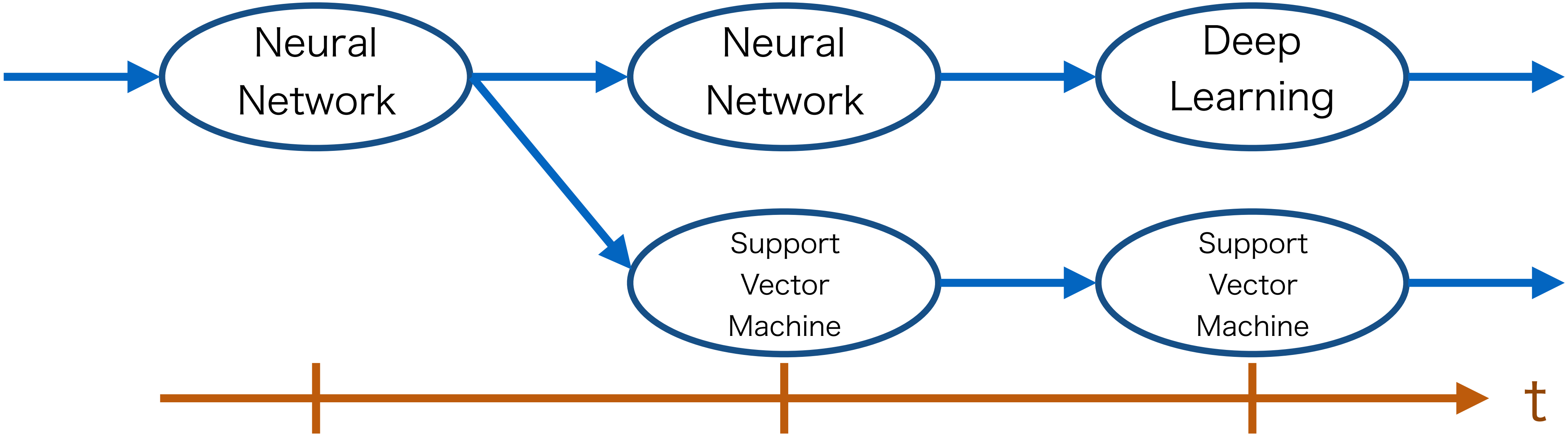
- Background
- Non-negative matrix factorization (NMF)
- Proposed method
- Experiments
- Conclusion and future work

Bibliographic DBs

- Google scholar, MS Academic Search, DBLP, CiteSeerX, ADS, Medline/PubMed, CiNii, …
- Huge academic information accumulated.
- Extract inherent academic knowledge to support researchers.

Topic evolution

- Changes of major research topics over time.



- Researchers can get:
  - major research topics,
  - how they evolved,
  - etc.

Related work (1/2)

- Topic detection
  - Probabilistic generative model
    - p-LSI [Hofmann. 1999]
    - LDA [Blei et al. 2003]
  - Matrix Factorization
    - Non-negative Matrix Factorization [Lee et al. 1999]
  - Graph analysis
    - term-graph [Jo et al. 2007]

- Non-negative matrix factorization (NMF) is attracting much attentions.
  - Lower computational cost than probabilistic model
  - High ability of topic detection as well as probabilistic models
  - Relatively simple algorithm

Related work (2/2)

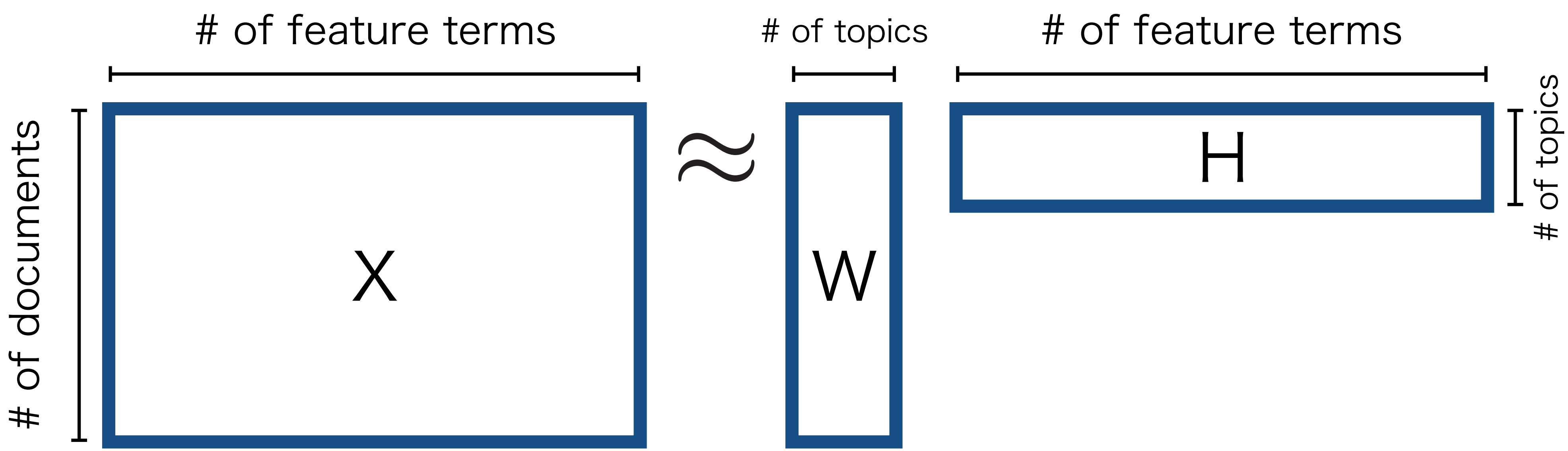
- Detecting topic evolutions
  - Using probabilistic model considering textual data and citations [He et al. 2009]
  - Using NMF introduced the topic transition matrix that explicitly connects past and present topics [Vaca et al. 2014]
  - Scheme of detecting a tendency of topic transitions in whole of the bibliographic database [Masada et al. 2012]
- Detecting community evolution
  - Scheme of detecting research community introduced new similarity measure of cluster similarity [Tajeuna et al. 2015]

Approach

- Partition DB by fixed-size time windows, and form doc-term matrices using title and abst.
- Detect topics in each matrix.
- Link similar topics in consecutive time windows.
- Exploit **CITATIONS** for better results.

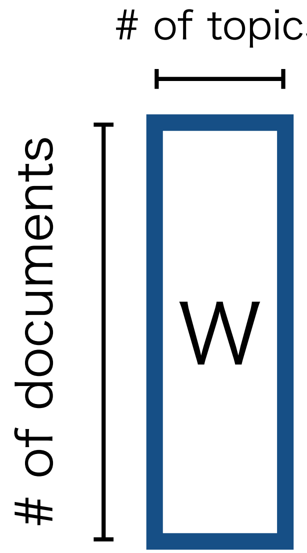
Applying NMF to doc-term matrix

- NMF is to approximate a non-negative matrix by two matrices with lower rank by optimizing loss function.
  - doc x term —> doc x topic * topic x term



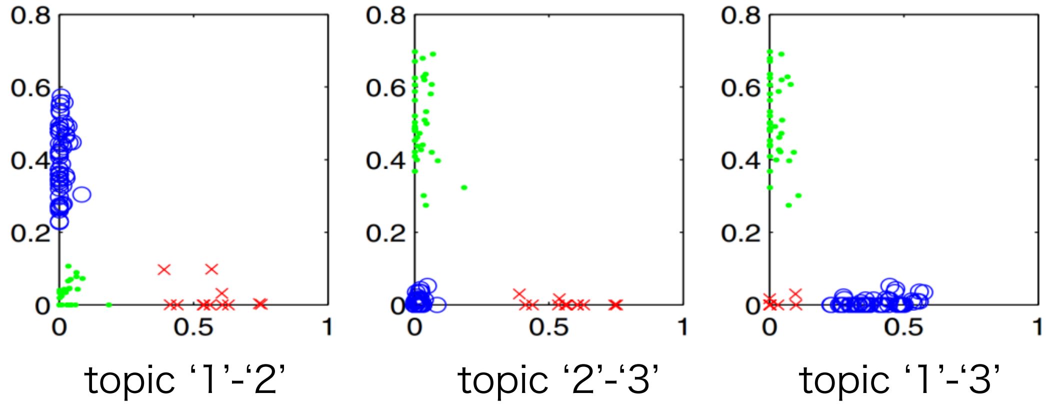
Matrix W: term distribution in each doc

- W : Ratio of each topic of each document contains



$$= \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_m \end{pmatrix}$$

Ex. Three topic vector bases

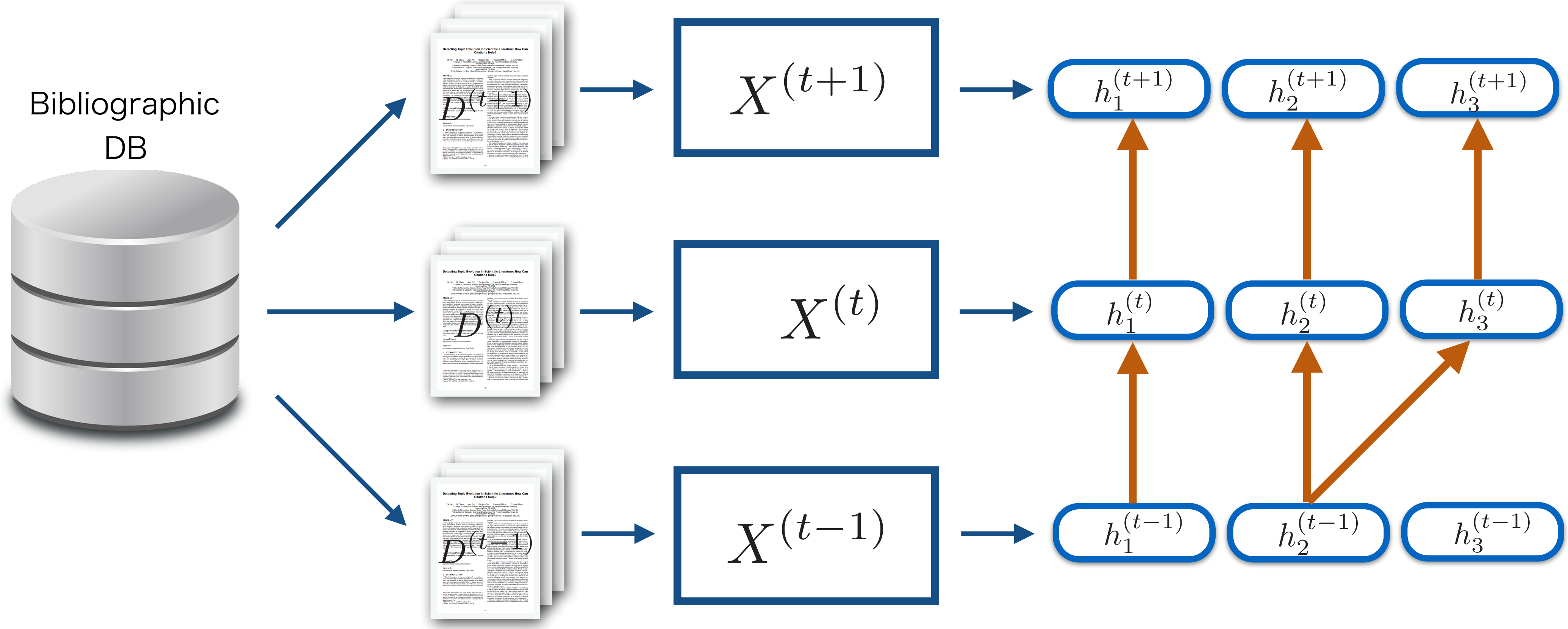


Source: W. Xu et al. "Document Clustering Based On Non-negative Matrix Factorization" In SIGIR'03

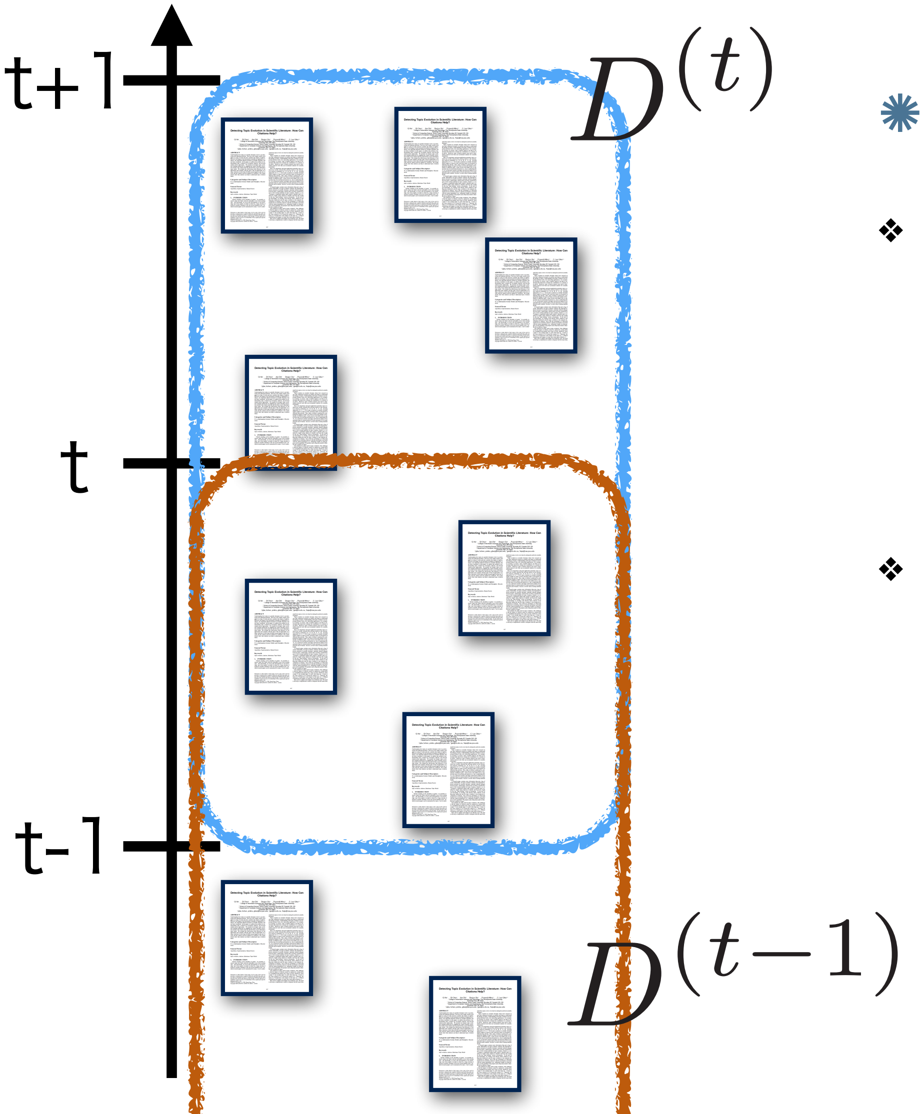
- Each doc is approximated with a topic vector of fewer dimensions.
- Topic-based cluttering of docs can be performed.

Proposal outline

Dataset partitioning → Transforming into matrices → Detecting topics using NMF → Connecting topics based on similarity

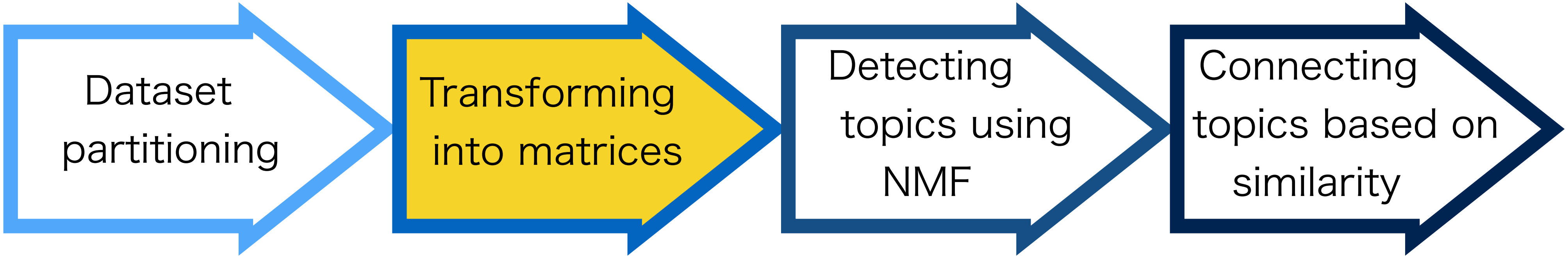


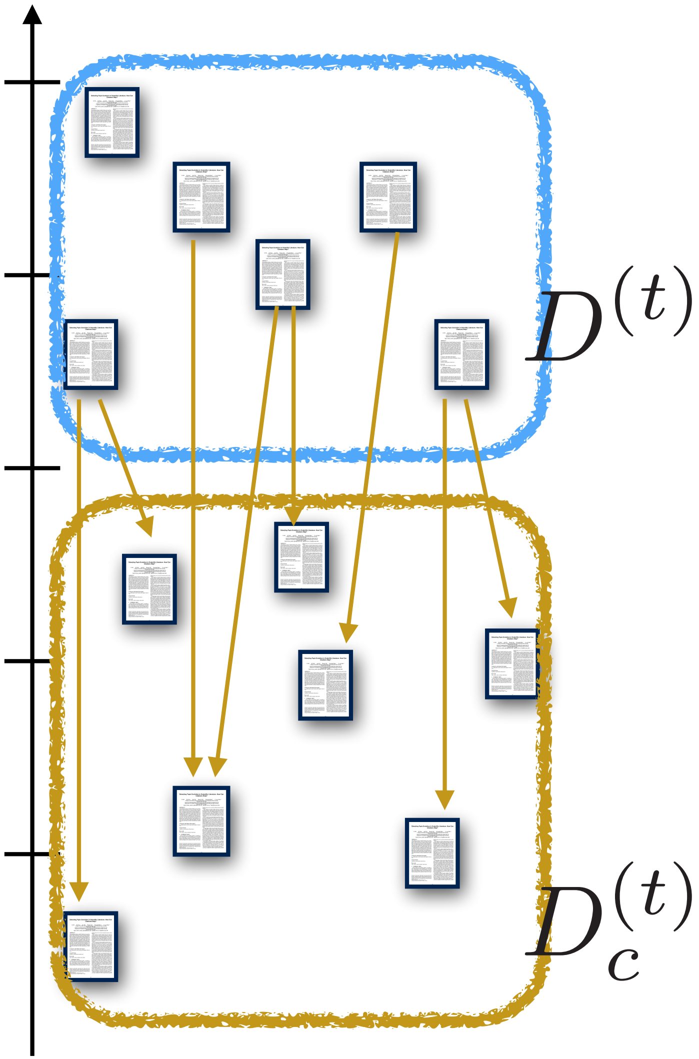
Dataset partitioning → Transforming into matrices → Detecting topics using NMF → Connecting topics based on similarity



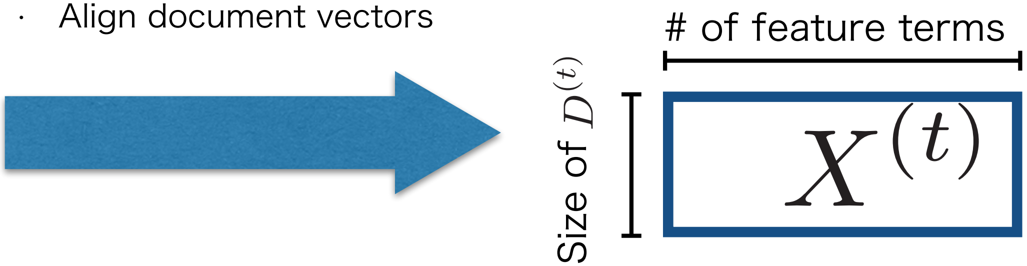
Dataset partitioning

- Overlap time intervals
  - To connect topics smoothly
  - To use as clue of connecting topics
- $D^{(t)}$: set of documents in time interval t

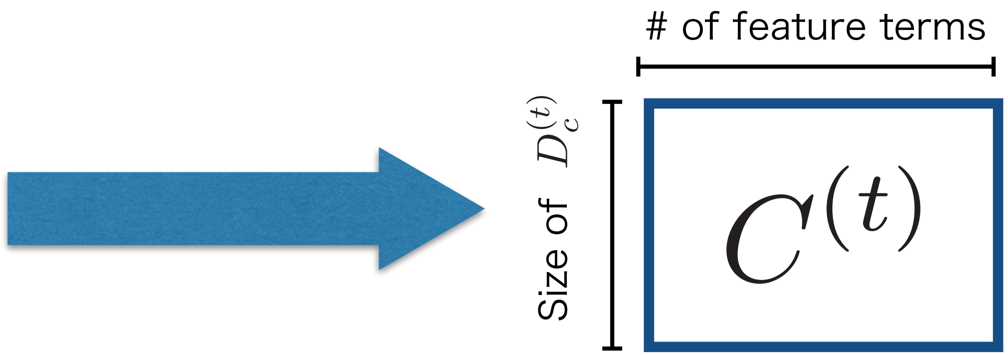




- Transform $D^{(t)}$ into matrix
  - Using title and abstract
  - Transform these textual data into vector (bag-of-words)
  - Align document vectors



- Transform documents that are cited by documents in time interval t into matrix
  - When detect topics, cited papers have important informations



Dataset partitioning → Transforming into matrices → **Detecting topics using NMF** → Connecting topics based on similarity

- Detect topics by applying NMF to matrix that is connected $X^{(t)}$ and $C^{(t)}$

$$\begin{bmatrix} X^{(t)} \\ C^{(t)} \end{bmatrix} \approx \begin{bmatrix} W_X^{(t)} \\ W_C^{(t)} \end{bmatrix} \begin{bmatrix} H^{(t)} \end{bmatrix}$$

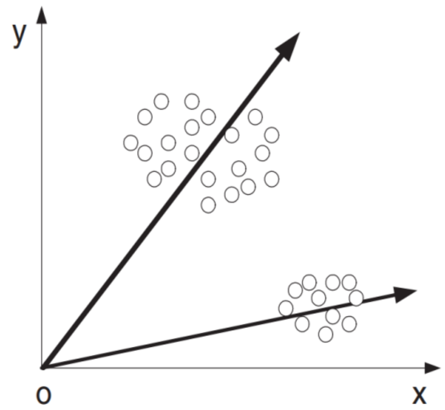
- Loss function:

$$L = \underset{W_X^{(t)}, W_C^{(t)}, H^{(t)}}{\arg\max} \left\| X^{(t)} - W_X^{(t)} H^{(t)} \right\|_F^2 + \delta \left\| C^{(t)} - W_C^{(t)} H^{(t)} \right\|_F^2$$

・ δ : Influence of cited papers for topics

Dataset partitioning → Transforming into matrices → Detecting topics using NMF → **Connecting topics based on similarity**

Linking similar topics



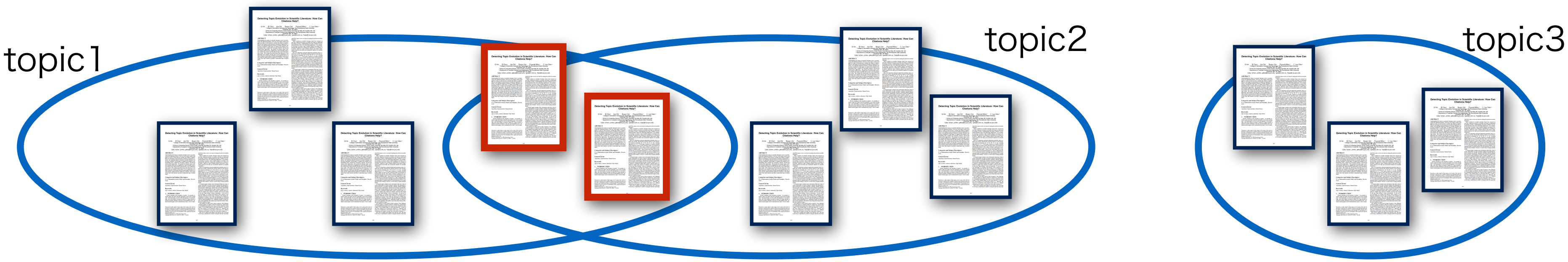
Source: W. Xu et al. "Document Clustering Based On Non-negative Matrix Factorization" In SIGIR'03

Approach 1

- Link topics if their word dist is similar.

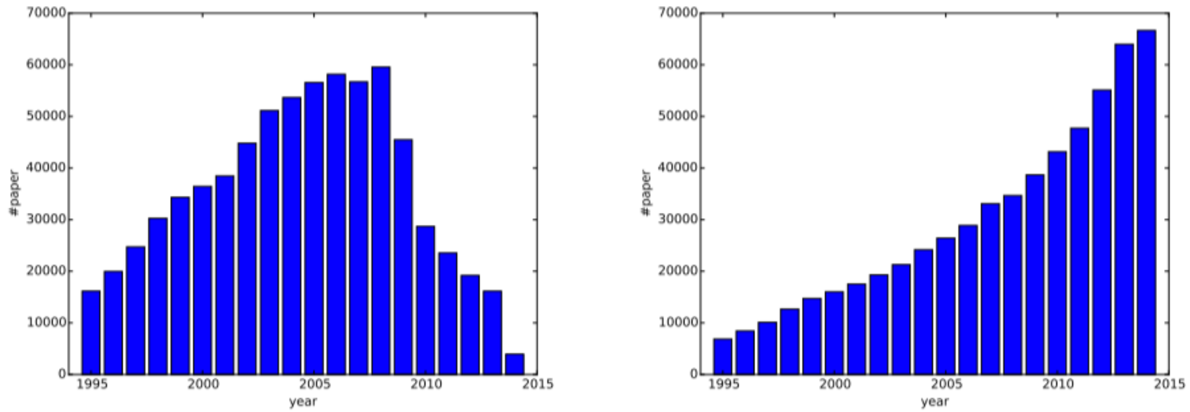
Approach 2

- Link topics if they share many docs in common.



Experiments

- Dataset :
  - CiteSeerX: 701,686 papers from 1996 to 2014.
  - arXiv: 945,889 papers from 1995 to 2014.

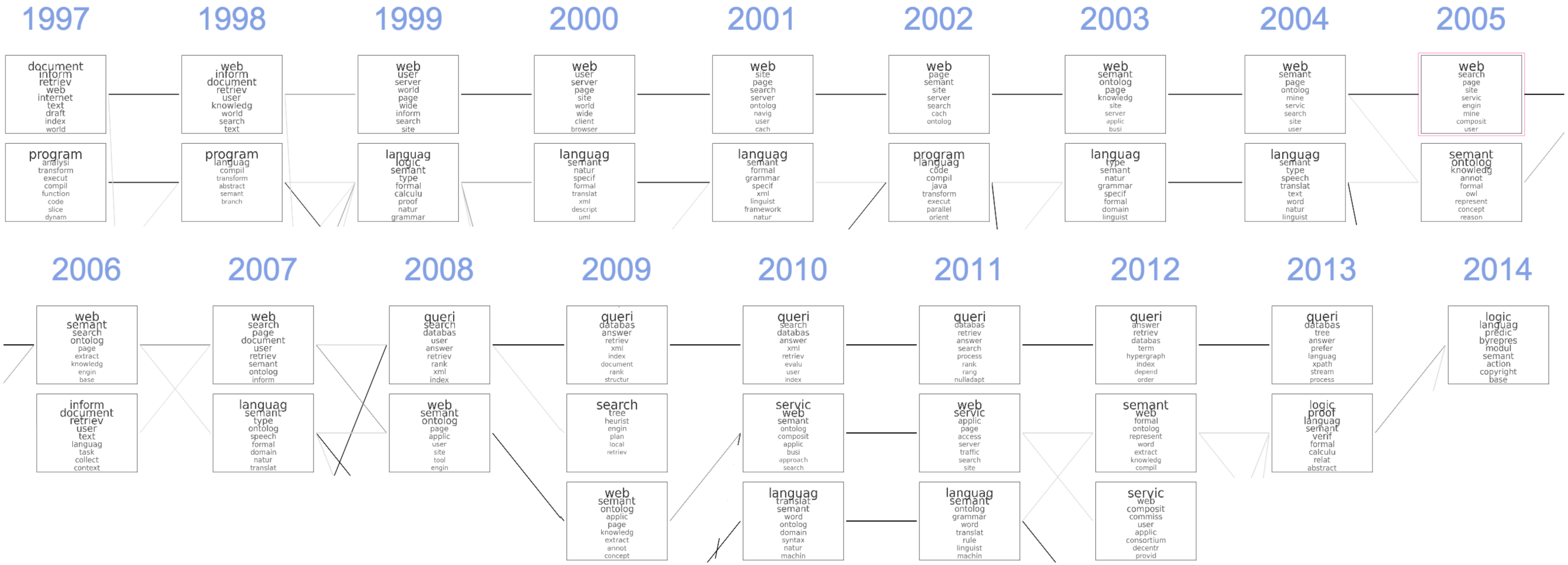


(a) Volume of CiteSeerX (b) Volume of arXiv

Figure 4. The data volume in each year

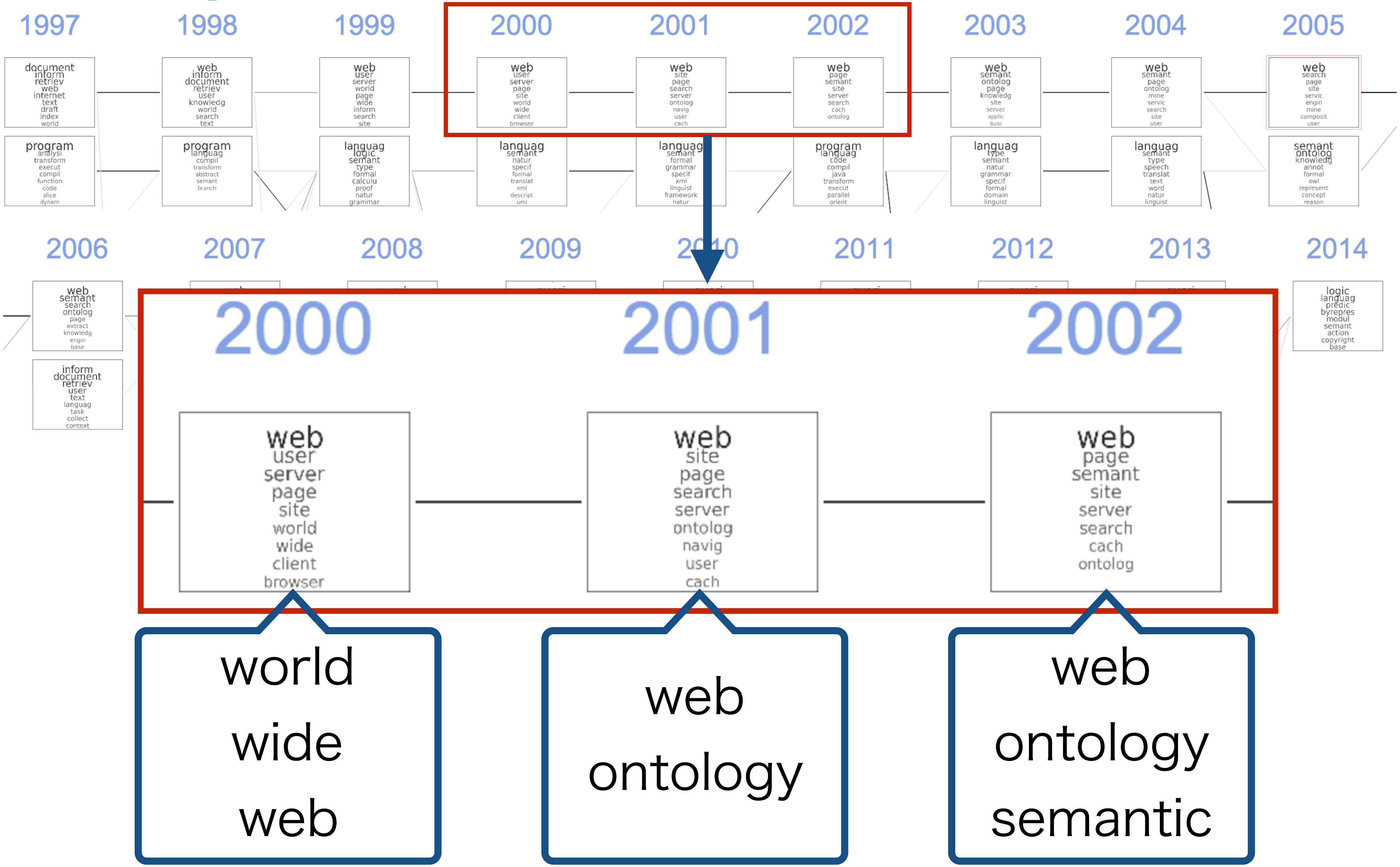
- Environment
  - Python 2.7 + numpy / spicy

Topic evolution: CiteSeerX

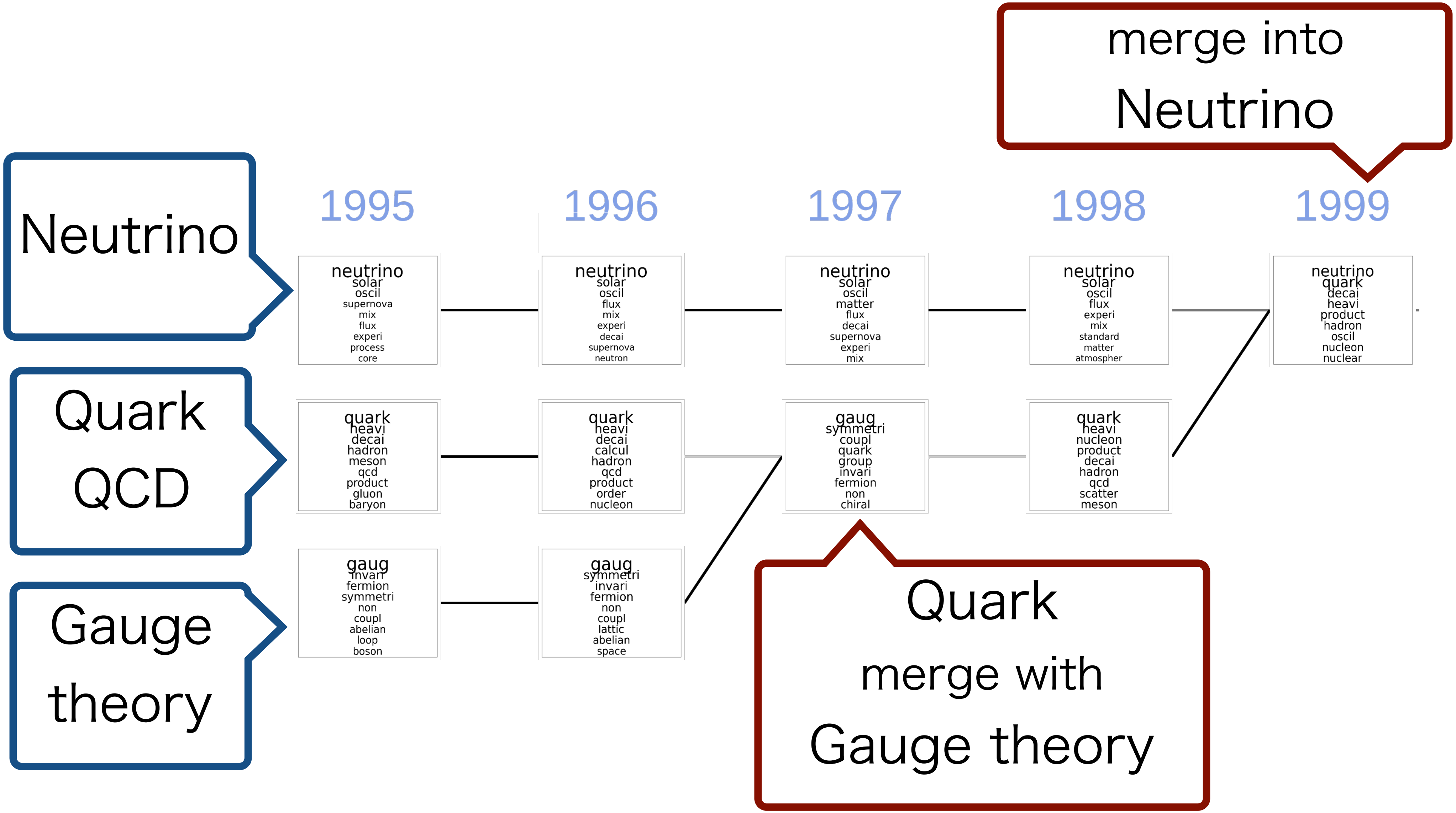


- A square represents topics detected in each year
- Terms in square describes each topic
- Size of terms indicates strength of term in topic
  - These are not labels by human tagging

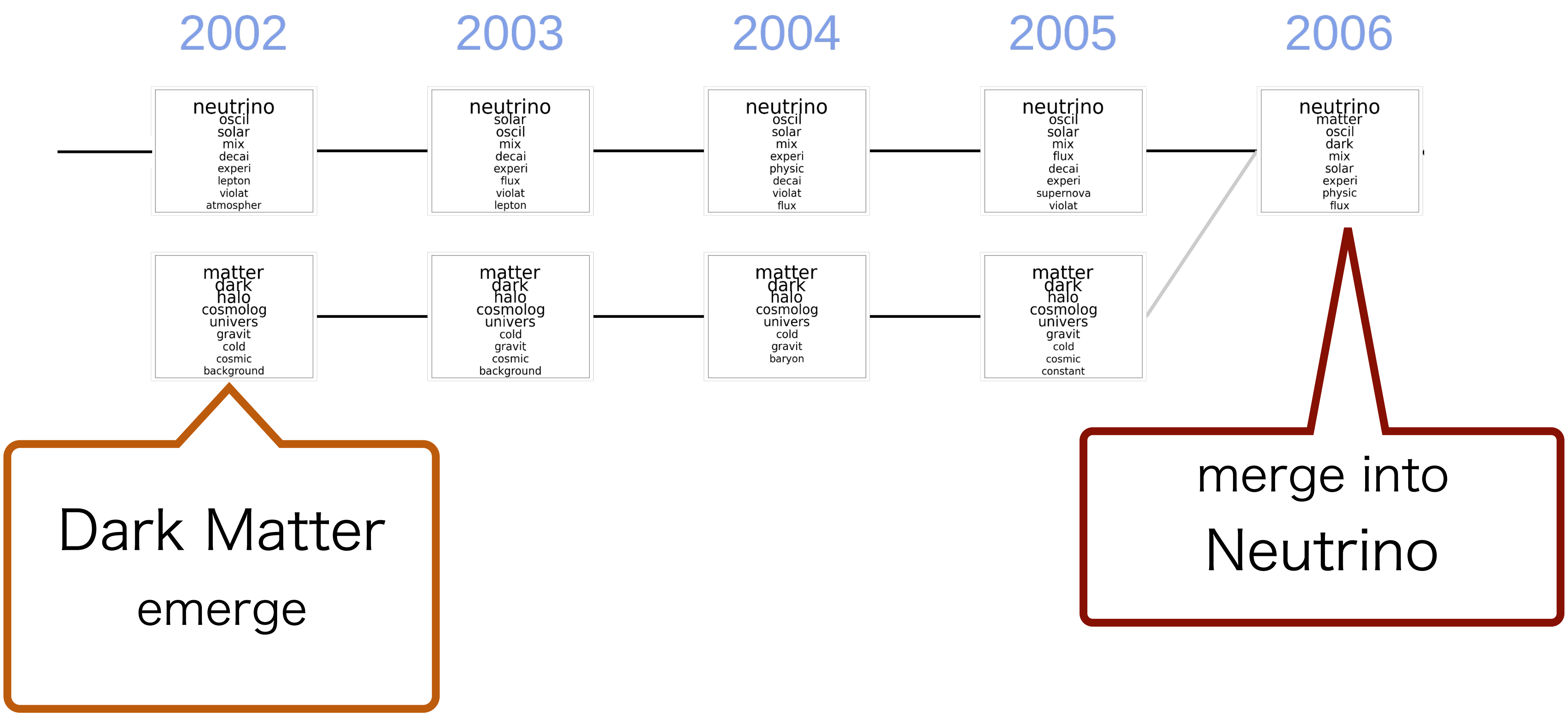
Topic evolution: CiteSeerX



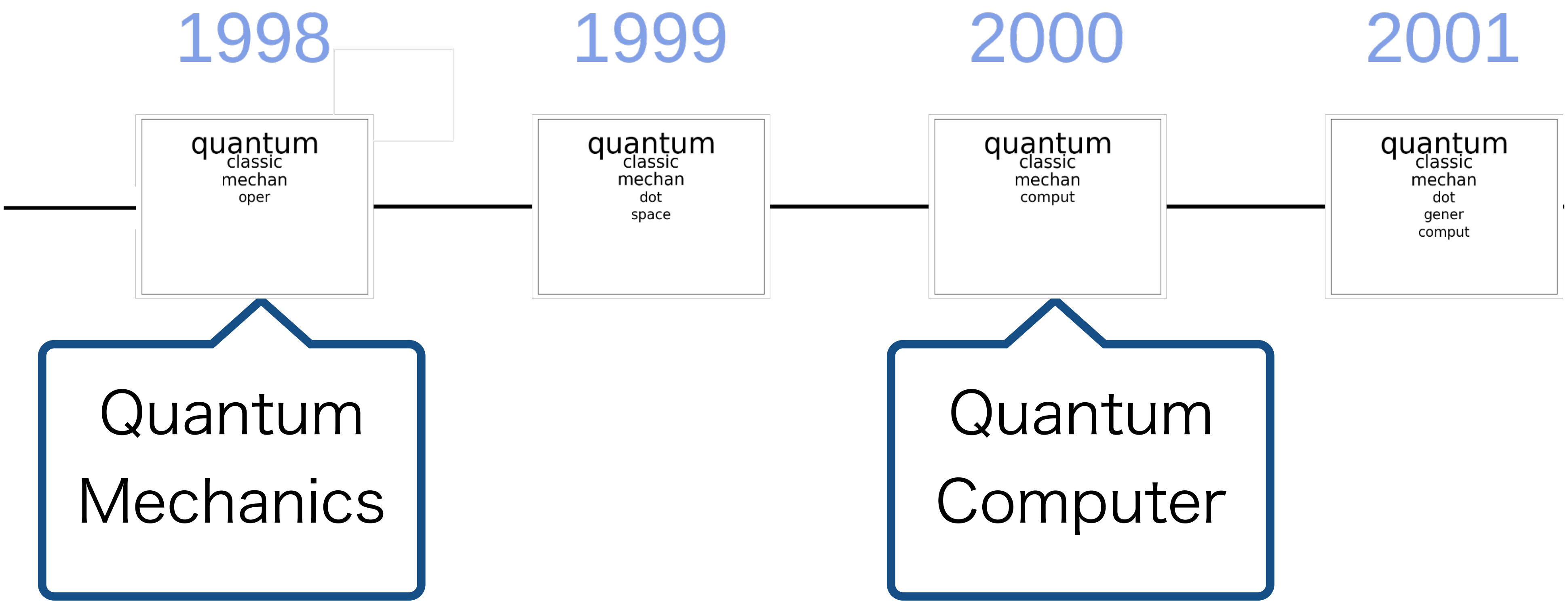
Topic evolution: arXiv (1/4)



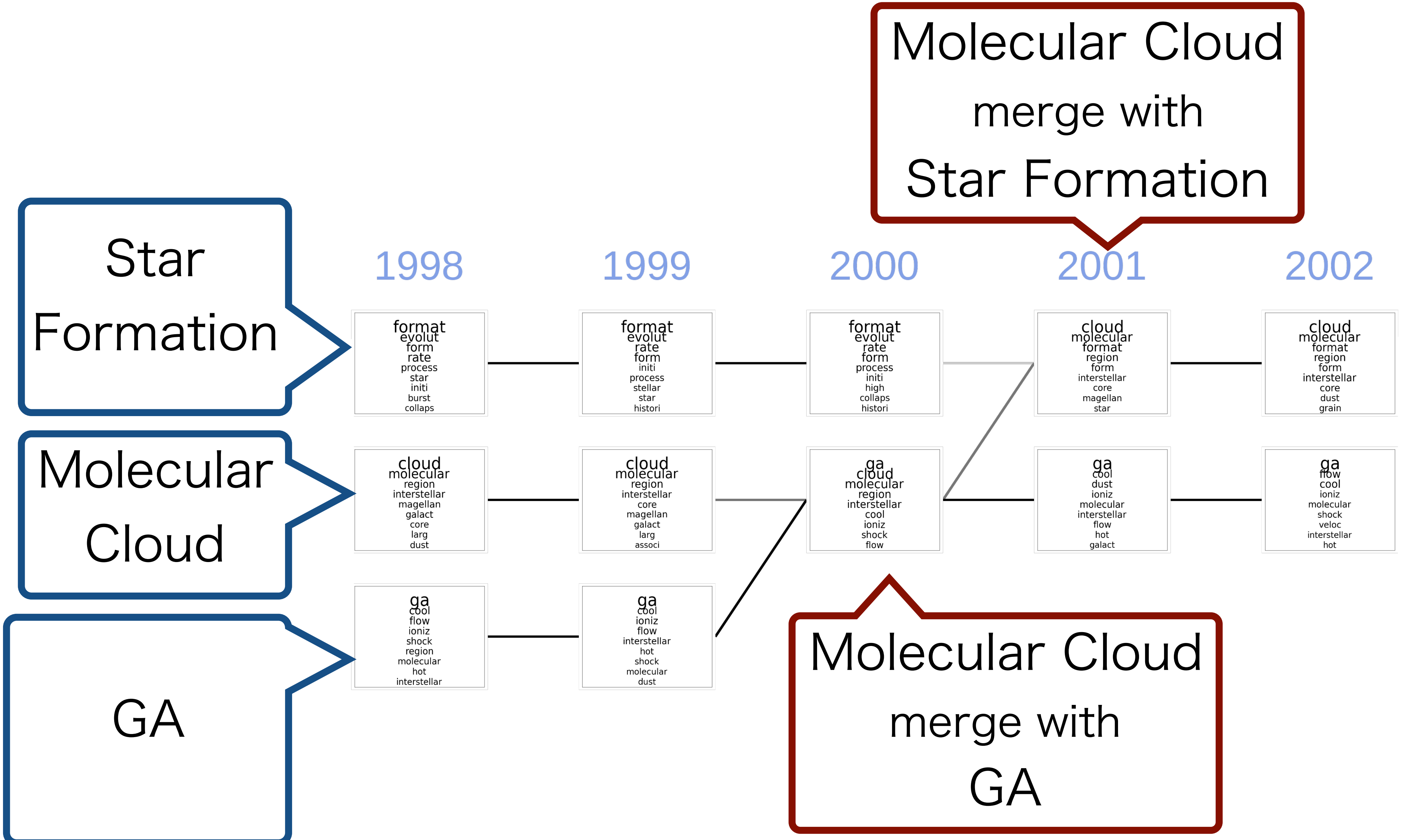
Topic evolution: arXiv (2/4)



Topic evolution: arXiv (3/4)



Topic evolution: arXiv (4/4)



Conclusion and Future Work

- Conclusion
  - We propose a scheme detecting topic evolution based on NMF exploiting citations
  - Our scheme successfully detect topic evolution
  - In a view point of diversity, our scheme greatly improve from a prior work
- Future work
  - Discuss about validity of topic and topic evolution
  - More efficient algorithms so that we can deal with large datasets