研究者から見た機関リポジトリ

天笠俊之
筑波大学計算科学研究センター
2016年6月23日（木）
機関リポジトリ新任担当者研修@NII

自己紹介

- 氏名：天笠俊之
- 所属：筑波大学計算科学研究センター
- 専門：データ工学
  - データベース
  - データマイニング
  - 科学分野へのデータベース応用
  - など

講演の内容

1. 機関リポジトリの意義
2. ビッグデータとしての学術リポジトリ
3. 学術リポジトリの分析
4. 学術リポジトリ分析の例
5. まとめ

研究者にとって論文とは

- 研究成果を表現・発表・蓄積するための主要な手段
  - 雑誌論文
    - オープンアクセスジャーナル
  - 国際会議論文
  - 研究会報告
  - テクニカルノート
  - 紀要
- 分野により位置づけが異なる
  - コンピュータサイエンス分野では，他分野に比べて国際会議論文の重要性が高い
  - 研究分野の進歩が早く，速報性が重視されるため（cf. プレプリントサーバ）

機関リポジトリの意義

- 教育・研究目的で生成された
  - 学術的・教育的資料
  - 得られた知見

  に対して，
  - 長期保管（アーカイブ）
  - 研究者・社会共有
  - 出版

  の手段／場を提供する.

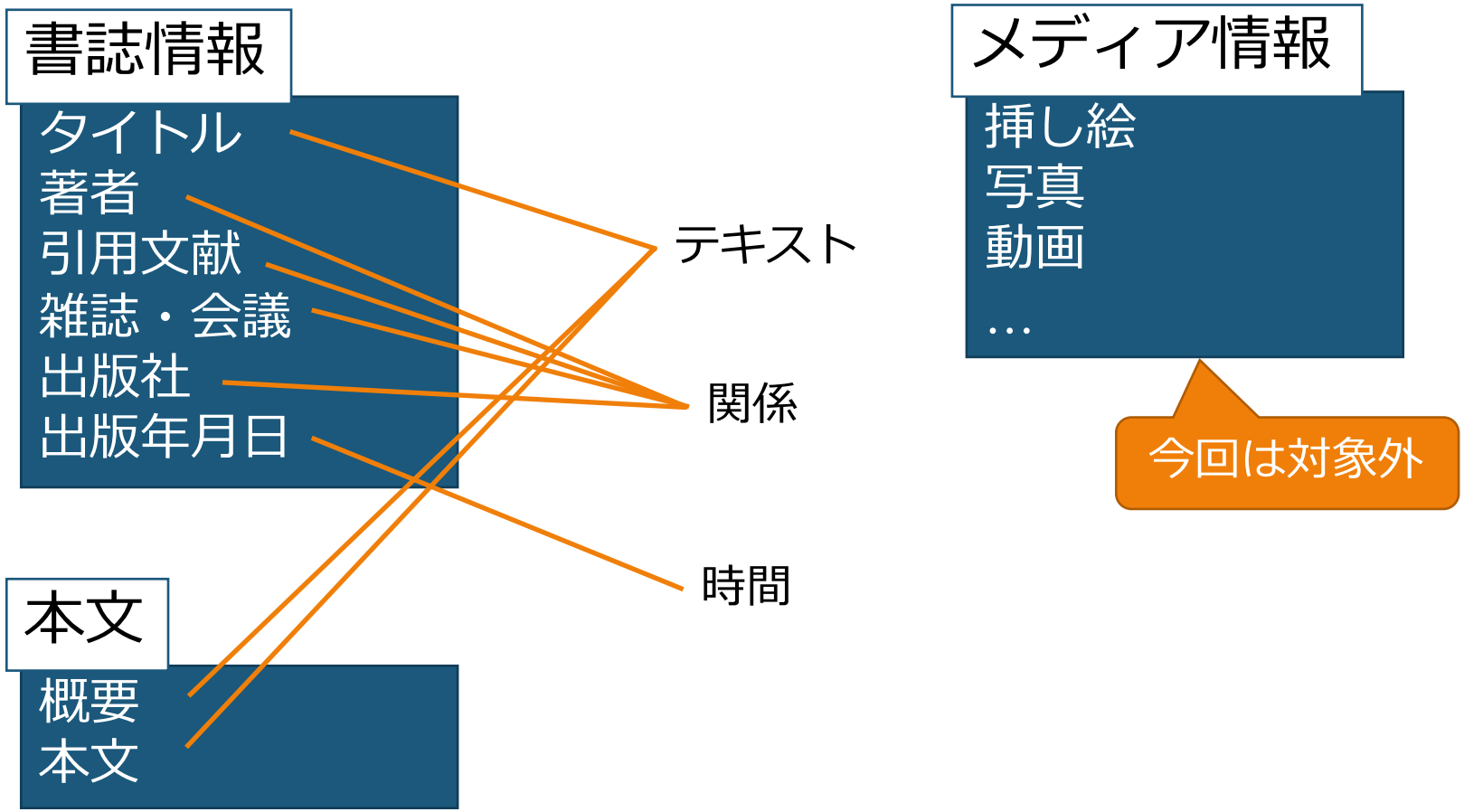
学術リポジトリの利用

- 文献検索の手段
- 文献入手の手段
- 文献出版の手段
- **情報資源（データセット）として**
  - 学術リポジトリそれ自身が貴重な情報源
  - さまざまな側面から分析が可能

学術リポジトリ ➔ **学術ビッグデータ**

ビッグデータとしての学術リポジトリ

- 学術リポジトリから得られる情報



学術リポジトリの多面的な分析

分析対象		分析の観点		
文献	×	テキスト	➔	研究分野・内容の分析
著者	×	関係	➔	研究者コミュニティの発見
文献	×	関係	➔	重要文献の検出
雑誌・会議	×	関係	➔	主要な学術雑誌の発見

- 複数の対象／観点を組み合わせた分析
- 時間軸を絡めた分析

学術リポジトリの分析は古くて新しい学問

- 古くからある学問
  - 科学計量学
  - 計量文献学
  - 引用・共引用分
  - インパクトファクタ（1955），h指数（2005）
- 近年の傾向・発展
  - 扱えるデータ量が膨大になった
  - 複雑なアルゴリズムが適用可能になった

  - 大規模テキストそのものの分析
  - 「研究者」や「コミュニティ」の分析
  - 研究トピックやコミュニティの時間変化

なぜ学術リポジトリを分析するのか

- 文献／研究者／雑誌・会議を定量的に評価したい
  - インパクトファクタ，h指数など
- 研究分野についての知識を（インスタンスから）発見したい
  - どのような研究が存在するか
  - 研究の間にどのような関係があるか
  - ある研究分野における主要な文献／研究者はどれか
- 研究活動をコミュニティの側面から分析したい
  - 研究グループの発見
  - 研究グループの相互の関係
  - コミュニティの出現／消滅の検出

学術リポジトリ分析

1. テキストデータ
2. 関係データ

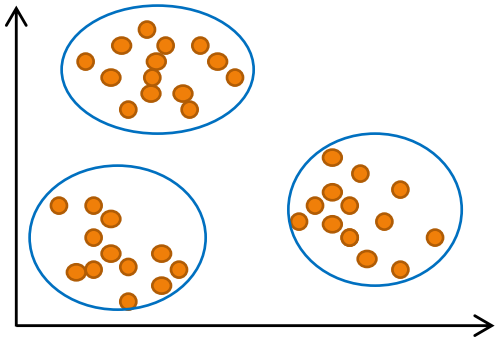
文献リポジトリからのデータ抽出：テキスト

- 文書ー単語行列
  - 行が文書，列が索引語に対応する行列
  - 各要素は出現頻度やTF-IDF法などで算出
  - 一般にサイズは巨大
    - 行数：文書数（～数十万）
    - 列数：索引語数（数万～数十万）

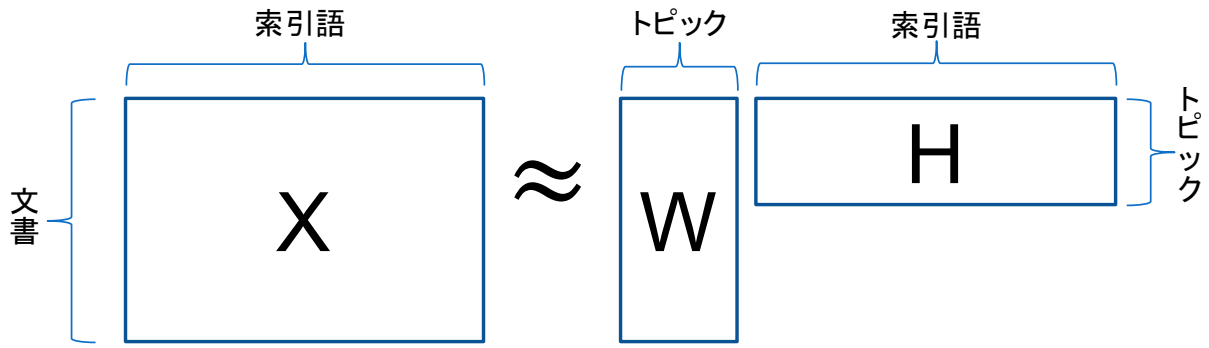
	単語1	単語2	単語3	単語4	単語5
文書1	3	4	2	0	0
文書2	0	2	3	0	1
文書3	1	0	1	4	3
文書4	0	0	0	5	3
文書5	5	6	3	0	1

テキストデータに対する典型的な分析処理

- クラスタ分析
  - 文献を（単語の出現傾向が）似たグループに分割
  - 例：k-means，凝集法など
- 行列の低ランク近似
  - 行列をよりサイズの小さい行列で近似（圧縮）
  - 似た使われ方をする単語がまとめられる
  - 例：LSI, pLSI, LDA, NMFなど



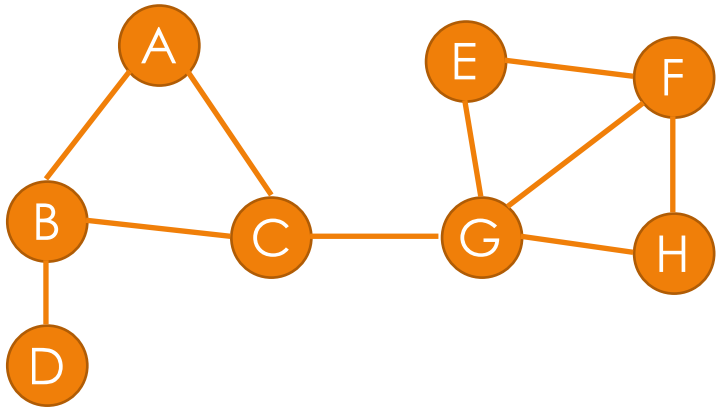
クラスタ分析



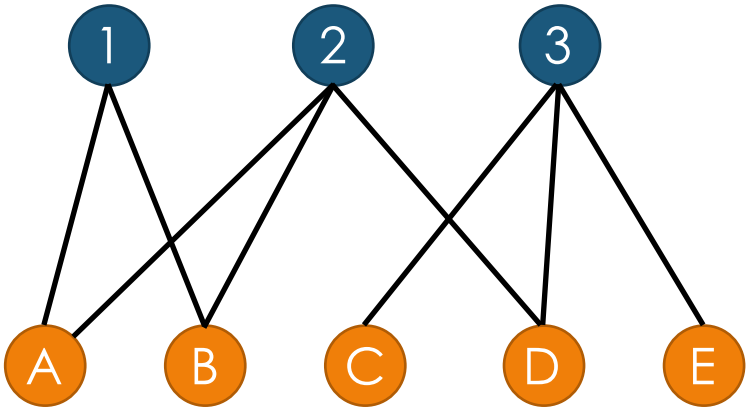
行列の低ランク近似

学術リポジトリからのデータ抽出：関係データ

- グラフ：対象間の関係を記述
  - 頂点：記述したい対象
  - 辺：頂点間に関係があることを記述
- 有向／無向グラフ，ラベル付きグラフ



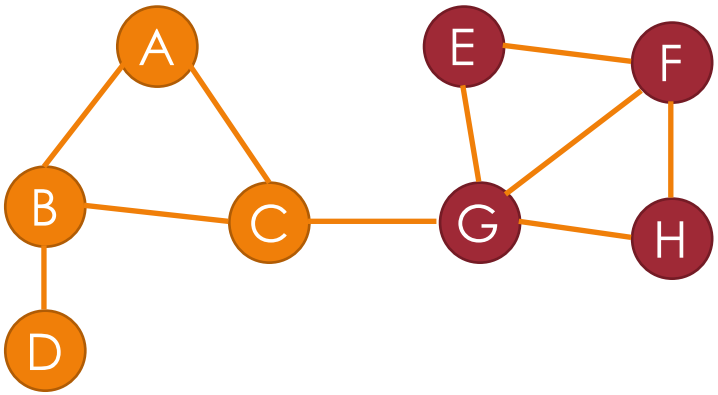
無向グラフ

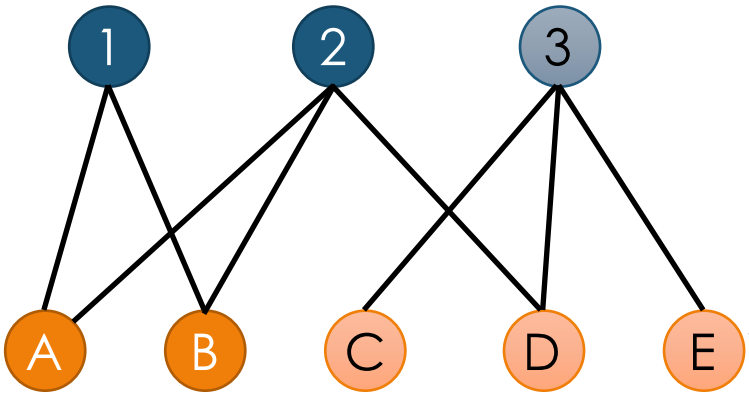


二部グラフ

グラフデータに対する典型的な分析処理

- リンク構造解析
  - 重要な頂点を発見
    - 重要な頂点からリンクされている頂点は重要
  - 例：PageRank, HITS, ObjectRankなど
- クラスタ分析
  - 辺が密に存在する部分グラフを発見
  - 関係が密に存在 ➔ コミュニティが存在
  - 例：モジュラリティクラスタリング，SCAN，ラベル伝搬法など





学術リポジトリ分析の例

RankClus [Y. Sun et al., EDBT 2009]

- 複数種類の頂点から構成されるグラフにおいて，ランキングを考慮した**クラスタリング**を行うアルゴリズムを提案
  - 質の高い研究者は質の高い雑誌／会議に論文を投稿
  - 質の高い雑誌／会議に採択される研究者は質が高い
- 学術リポジトリ（DBLP bibliography）に適用

Table 5: Top-10 Conferences in 5 Clusters Using RankClus

	DB	Network	AI	Theory	IR
1	VLDB	INFOCOM	AAMAS	SODA	SIGIR
2	ICDE	SIGMETRICS	IJCAI	STOC	ACM Multimedia
3	SIGMOD	ICNP	AAAI	FOCS	CIKM
4	KDD	SIGCOMM	Agents	ICALP	TREC
5	ICDM	MOBICOM	AAAI/IAAI	CCC	JCDL
6	EDBT	ICDCS	ECAI	SPAA	CLEF
7	DASFAA	NETWORKING	RoboCup	PODC	WWW
8	PODS	MobiHoc	IAT	CRYPTO	ECDL
9	SSDBM	ISCC	ICMAS	APPROX-RANDOM	ECIR
10	SDM	SenSys	CP	EUROCRYPT	CIVR

Yizhou Sun, Jiawei Han, Peixiang Zhao, Zhijun Yin, Hong Cheng, Tianyi Wu:
RankClus: integrating clustering with ranking for heterogeneous information network analysis. EDBT 2009: 565-576

引用情報を利用した文書データベースからのトピック変遷検出

伊藤寛祥*，天笠俊之**，北川博之**

* 筑波大学大学院システム情報工学研究科

** 筑波大学計算科学研究センター

The 26th Int' Conf. on Information Modelling and Knowledge Bases (EJC 2016), Tampere, Finland, June 6-10, 2016.

研究背景

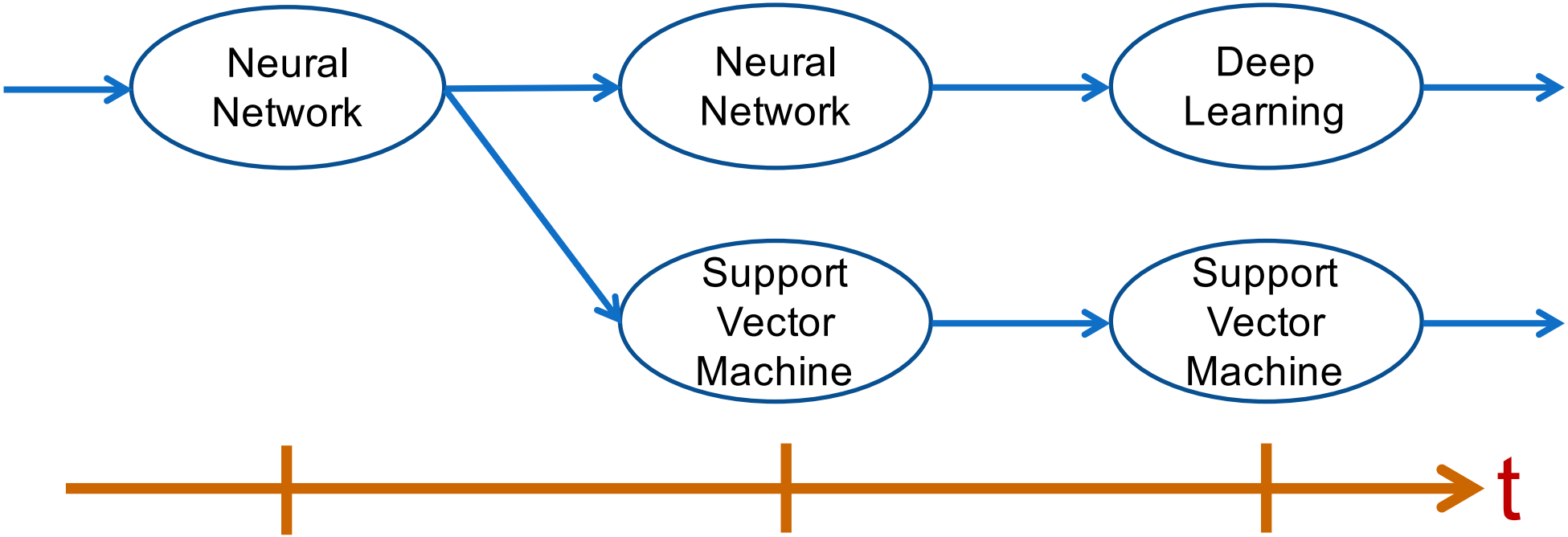
- 近年、学術分野において論文データのリポジトリの整備が進んでいる
  - DBLP : コンピュータサイエンス
  - PUBMED,MEDLINE： 医学
  - ADS, arXiv : 宇宙物理学

論文データベースに対する分析手法は注目を集めている

- 主要論文の抽出
- 研究者ネットワークの抽出
- 論文データベースにおけるトピック変遷の検出

トピック変遷

- 時間経過によるトピックの変化



◆研究者に有効であるさまざまな情報を得る手掛かりとなる

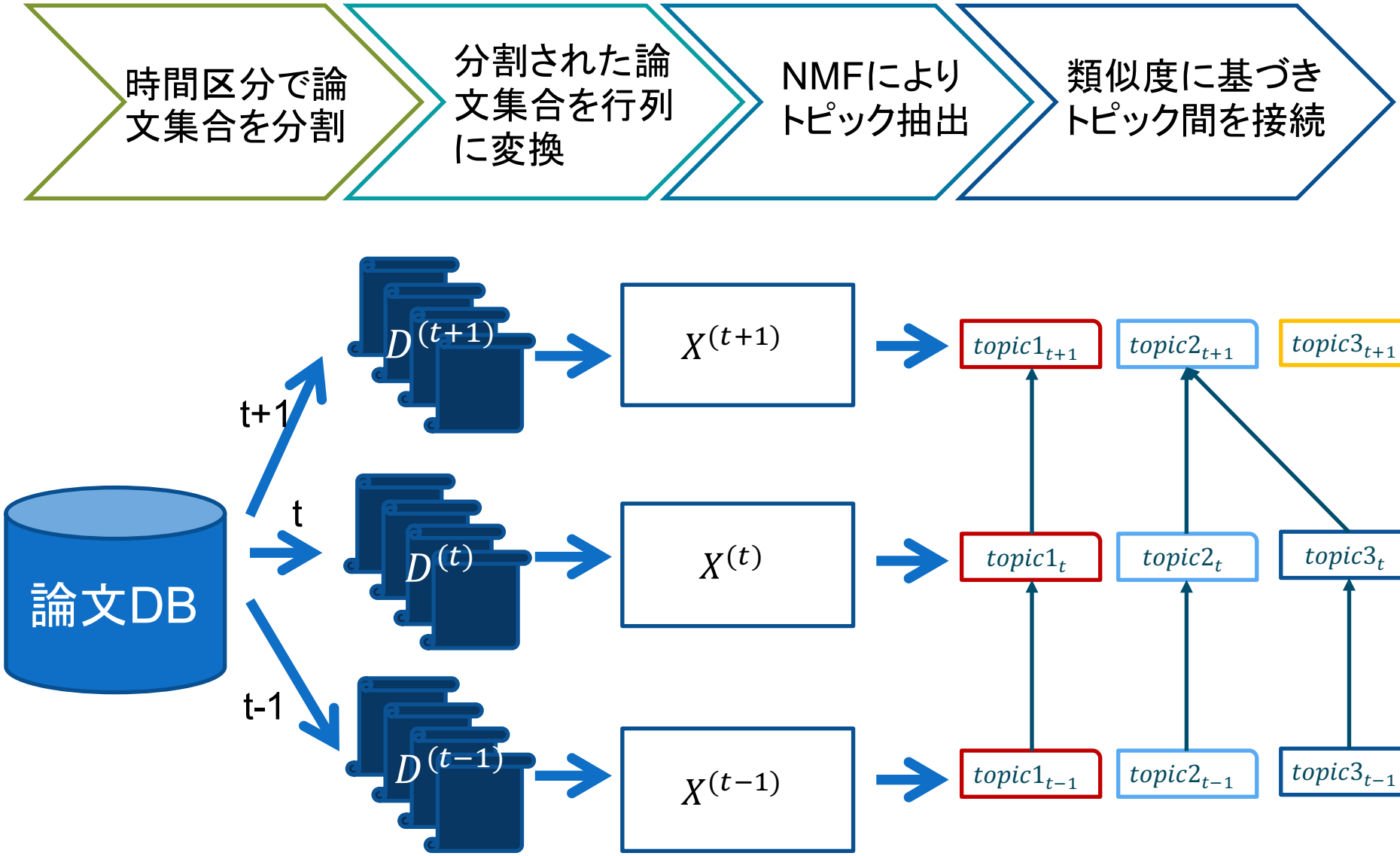
- 自分の研究がどのトピックの系統に属するか
- トピックの起源となった論文は何であったか
- etc

研究の目的

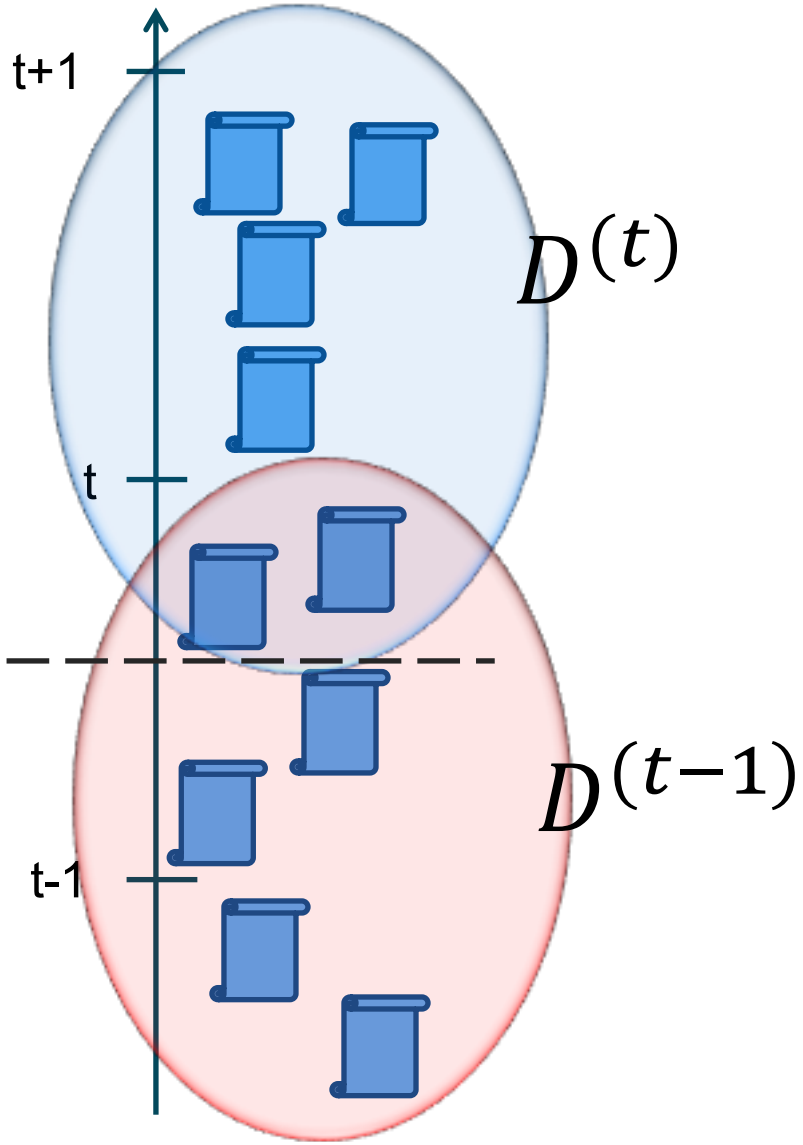
- 文献データベースから主要な研究トピックの変遷を抽出

- アプローチ
  - 文献の引用情報を活用
  - 非負値行列分解（non-negative matrix factorization; NMF）を利用

提案手法の概略


時間区分で論文集合を分割
分割された論文集合を行列に変換
NMFによりトピック抽出
類似度に基づきトピック間を接続
論文DB
t+1
t
t-1
$D^{(t+1)}$
$D^{(t)}$
$D^{(t-1)}$
$X^{(t+1)}$
$X^{(t)}$
$X^{(t-1)}$
$topic1_{t+1}$
$topic2_{t+1}$
$topic3_{t+1}$
$topic1_t$
$topic2_t$
$topic3_t$
$topic1_{t-1}$
$topic2_{t-1}$
$topic3_{t-1}$

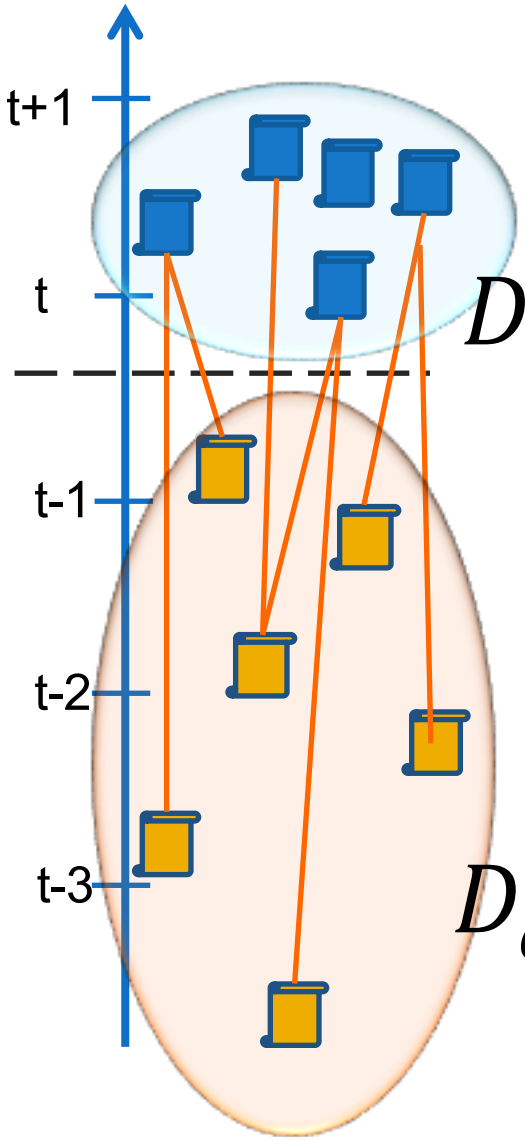
時間区分で論文集合を分割 → 分割された論文集合を行列に変換 → NMFによりトピック抽出 → 類似度に基づきトピック間を接続



時間区分

- 期間をオーバーラップさせる
  - 隣接する時間区分におけるトピック間が滑らかに接続されるように
  - 隣接する時間区分におけるトピック間を接続するときの手掛かりにする
- その期間 t における論文集合を$D^{(t)}$とする

時間区分で論文集合を分割 → 分割された論文集合を行列に変換 → NMFによりトピック抽出 → 類似度に基づきトピック間を接続



- 時間区分 t における論文集合のタイトル、アブストラクトをBag-of-wordsに変換した後、縦に並べることで行列に変換

$D^{(t)}$ → $X^{(t)}$

$D^{(t)}$ の文書数 × 特徴単語数

- 論文からトピックを抽出する際に被引用論文は有効な情報になる
  - 時間区分 t における論文が引用している論文も$X^{(t)}$とは別に行列に変換する

$D_c^{(t)}$ → $C^{(t)}$

$D_c^{(t)}$ の文書数 × 特徴単語数

時間区分で論文集合を分割 → 分割された論文集合を行列に変換 → NMFによりトピック抽出 → 類似度に基づきトピック間を接続

- 論文行列 $X^{(t)}$ と被引用論文行列 $C^{(t)}$ を結合させた行列にNMFを適用してトピックを抽出する

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

- 以下の損失関数を最適化することで行列分解

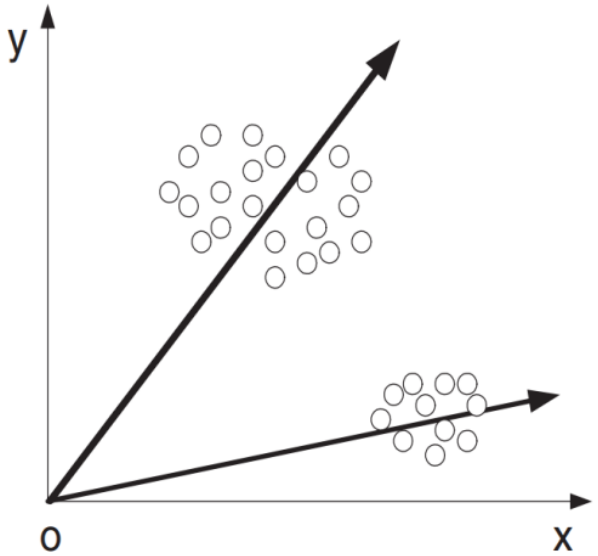
$$L = arg \min_{W_X^{(t)}, W_C^{(t)}, H^{(t)}} \left\| X^{(t)} - W_X^{(t)} H^{(t)} \right\|_F^2 + \delta \left\| C^{(t)} - W_C^{(t)} H^{(t)} \right\|_F^2$$

  - δ ： 被引用論文がトピックに与える影響力

時間区分で論文集合を分割 → 分割された論文集合を行列に変換 → NMFによりトピック抽出 → 類似度に基づきトピック間を接続

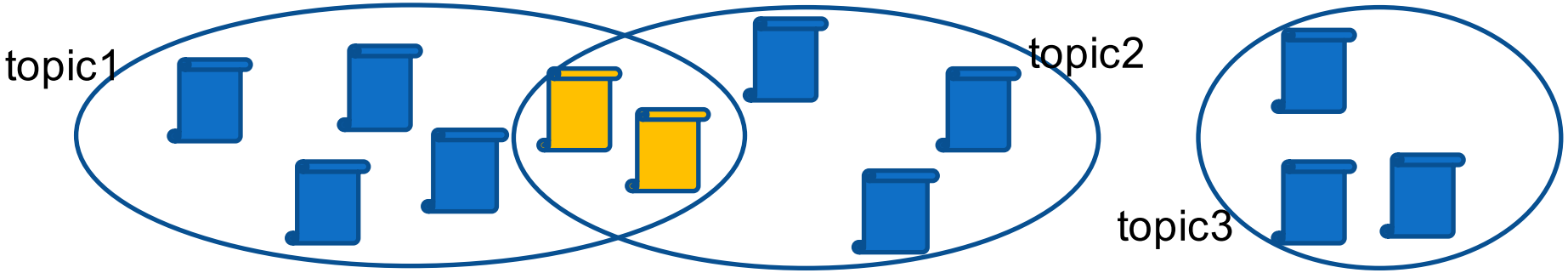
◆手法1

◆トピックの内容（単語の出現傾向）が似ていたら類似トピックと判定.



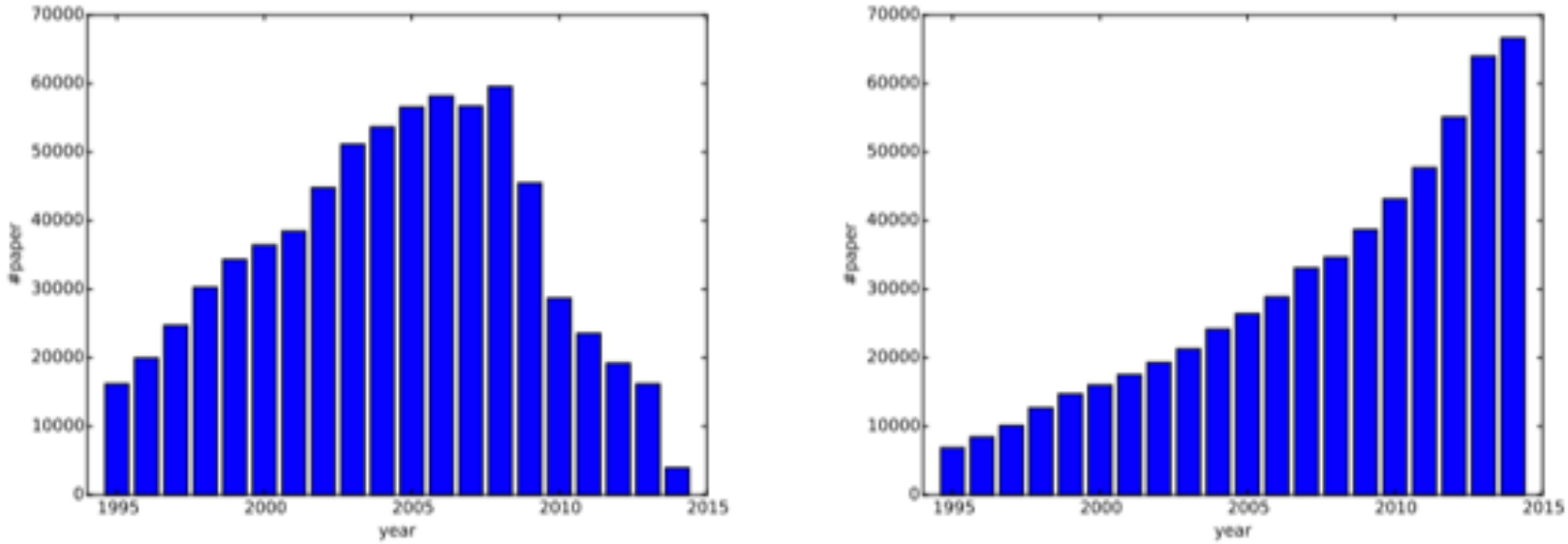
◆手法2

◆多くの論文が共通していたら類似トピックと判定.



評価実験

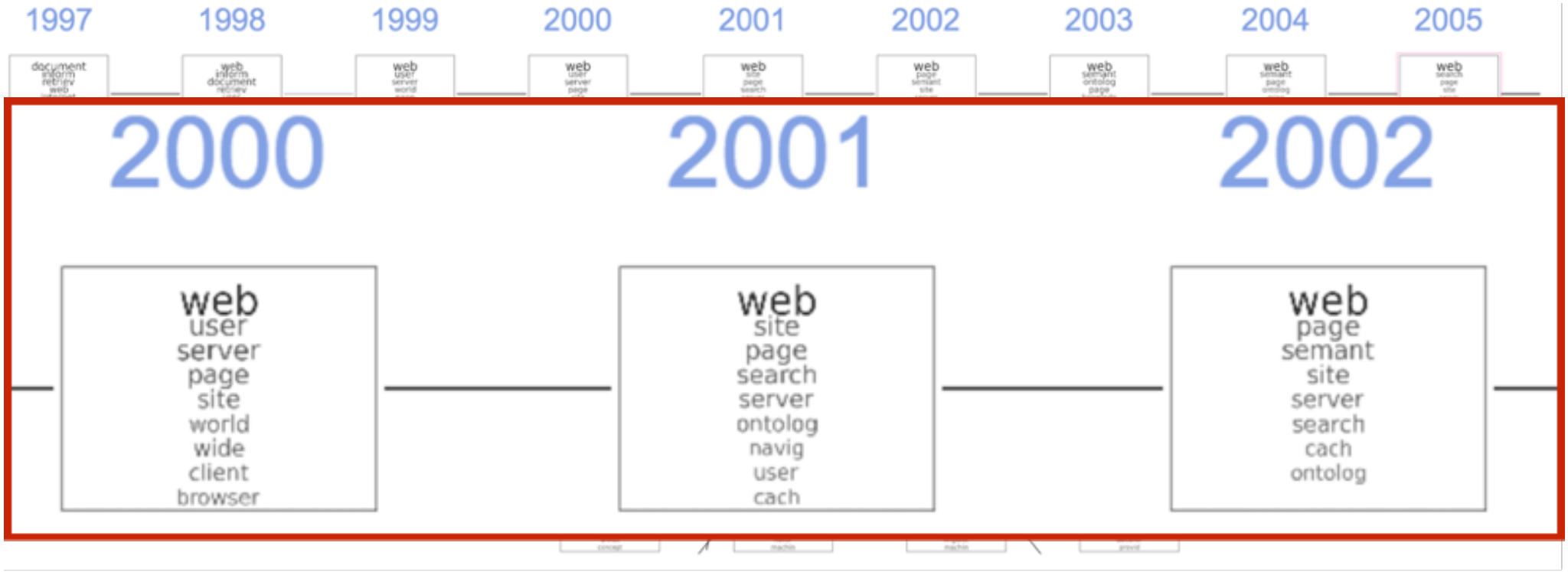
- CiteSeerX: 701,686件（1996-2014）
- arXiv: 945,889件（1995-2014）



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

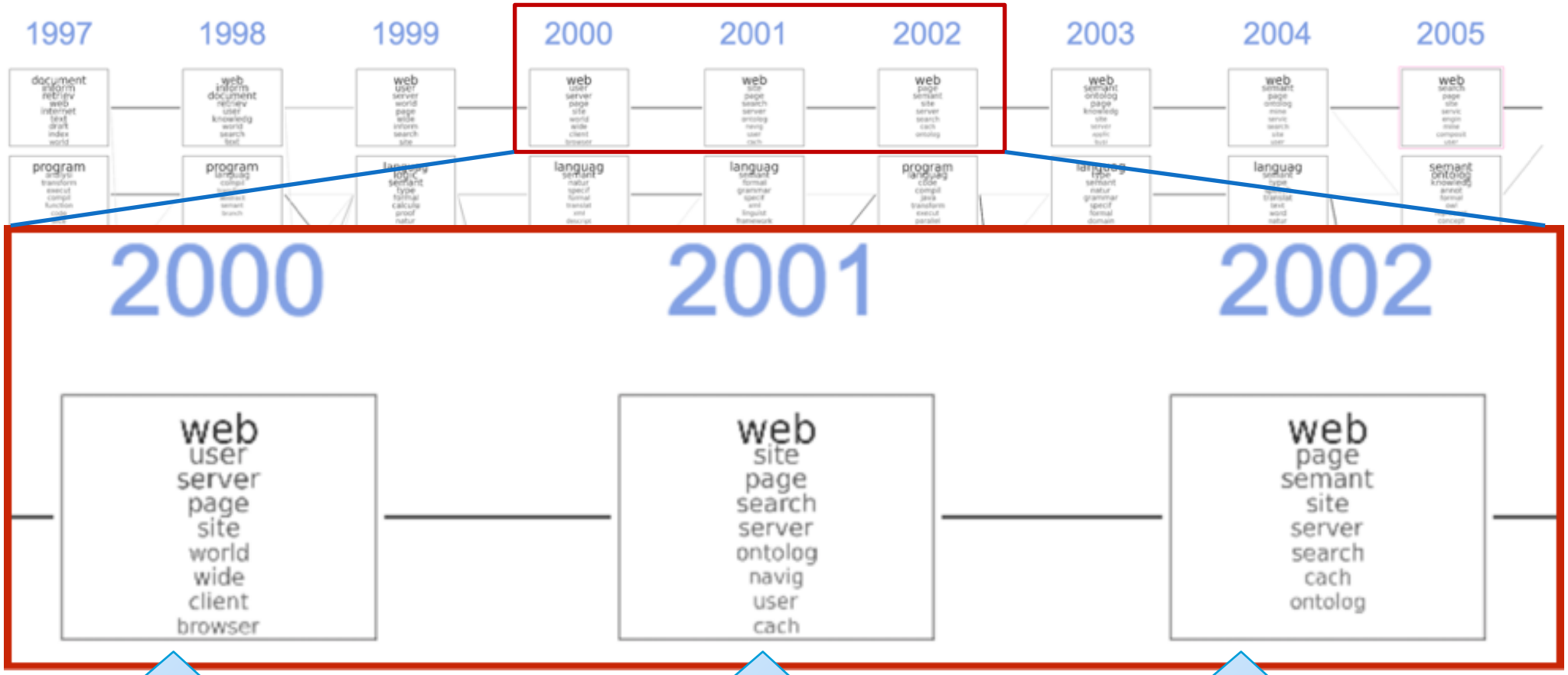
Figure 4. The data volume in each year

検出されたトピック変遷（CiteSeerX）

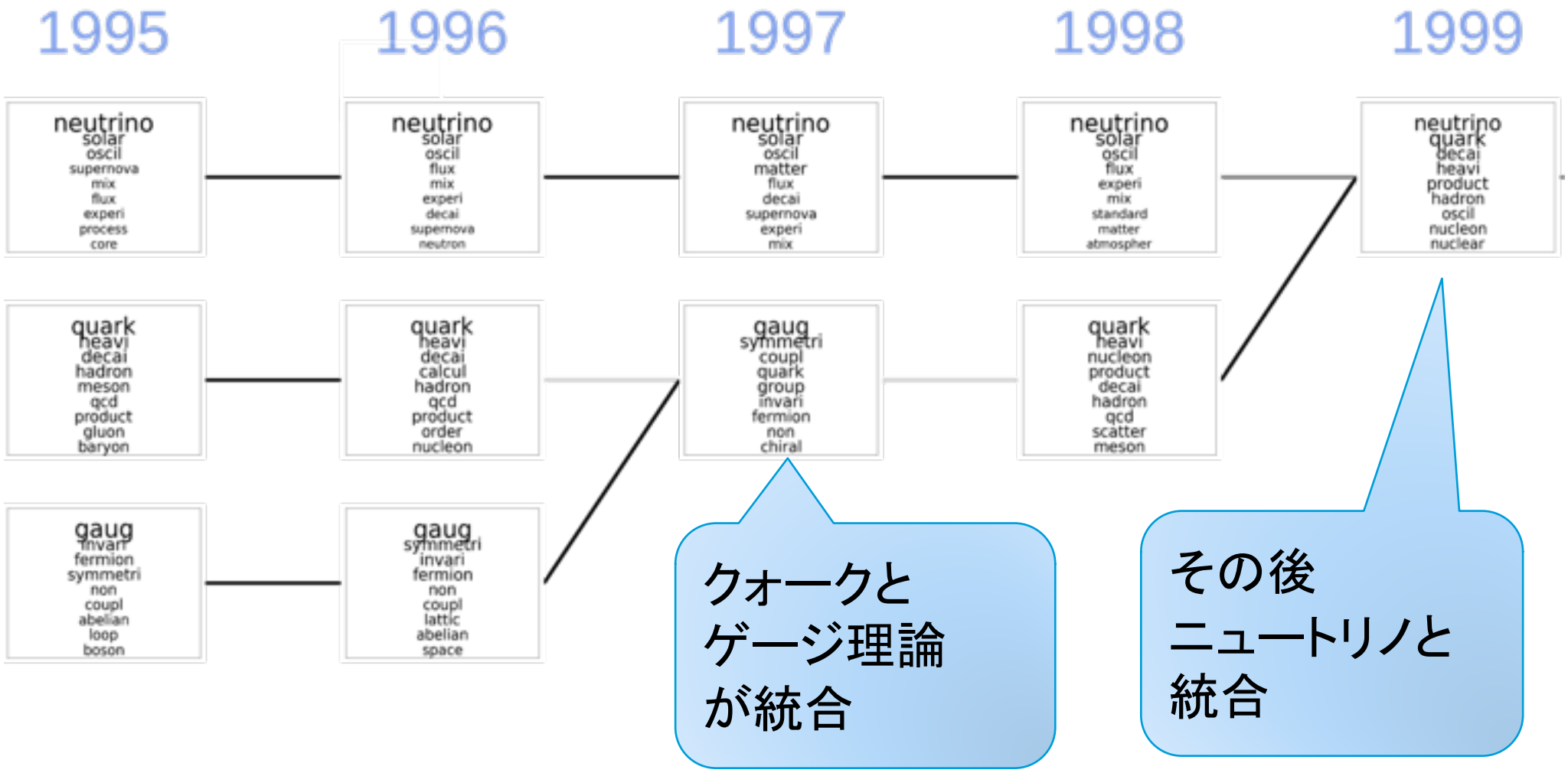


- 矩形がトピックを表す.
- 矩形中の単語は，トピックにおいて寄与の大きい単語を表す.
- 関連するトピックが接続されている.

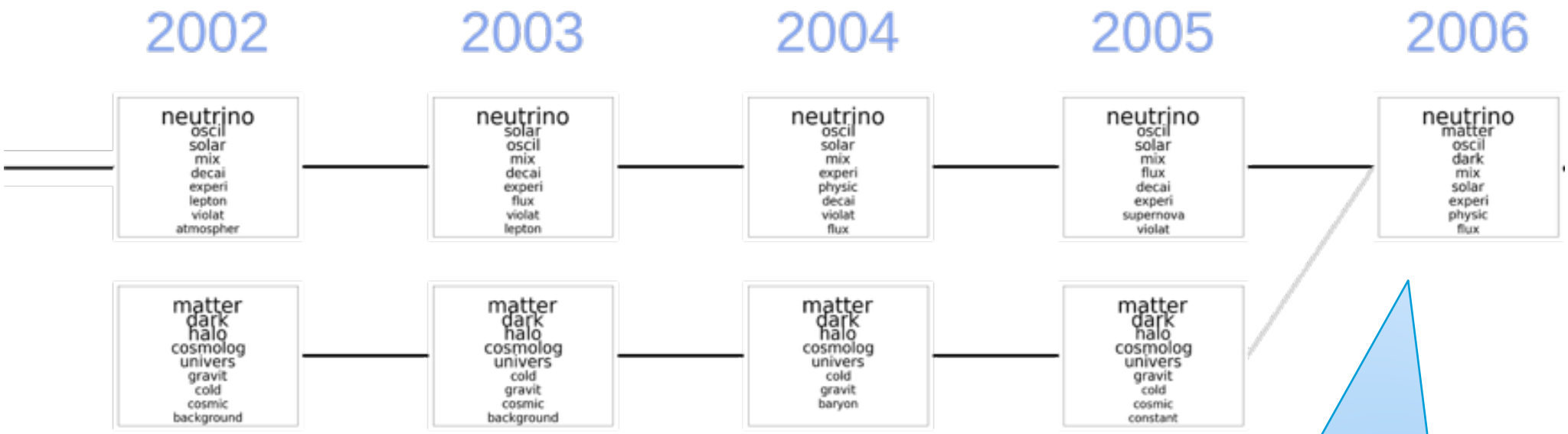
検出されたトピック変遷（CiteSeerX）



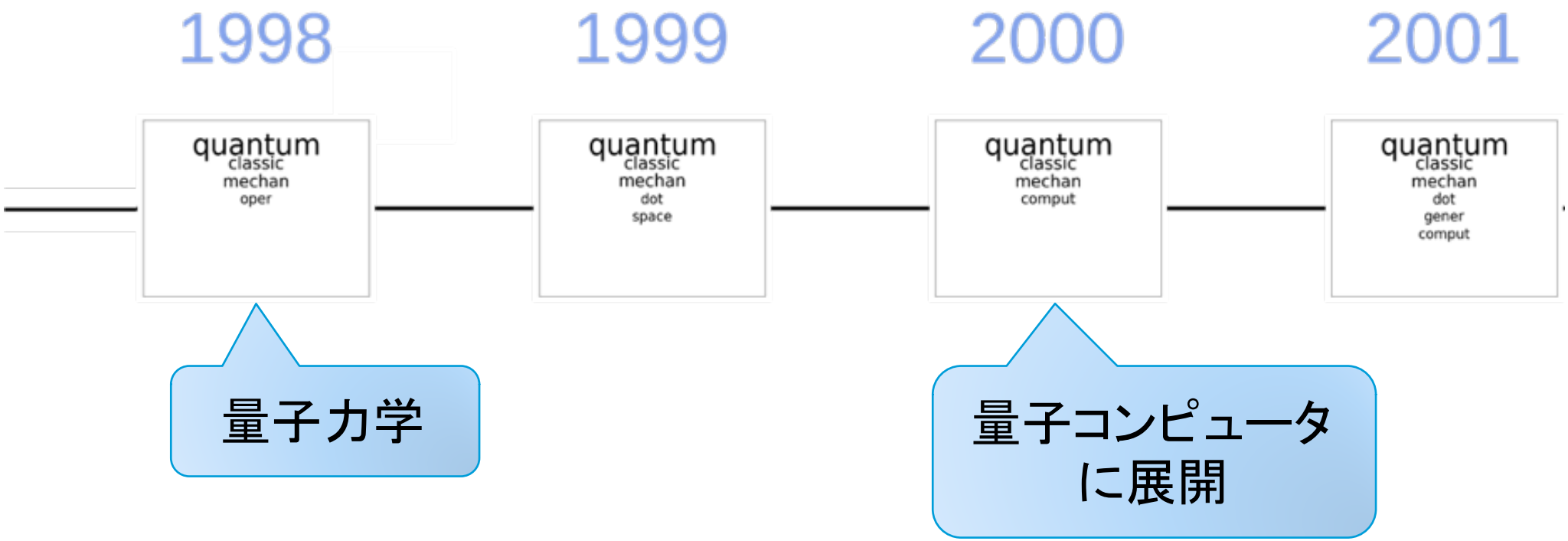
検出されたトピック変遷（arXiv）



検出されたトピック変遷（arXiv）



検出されたトピック変遷（arXiv）



まとめ

- コンピュータサイエンス研究者から見た学術リポジトリ
  - 学術リポジトリから得られる情報
  - 主な分析手法
  - 学術リポジトリ分析の例