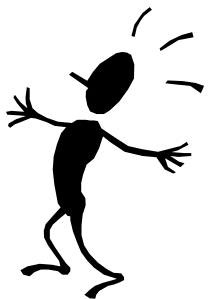
Colored Bottom-up DataGuide による半構造データの差異発見のためのスキーマ生成･可視化機構

神戸大学大学院自然科学研究科
小島岳史
清光英成
田中克己

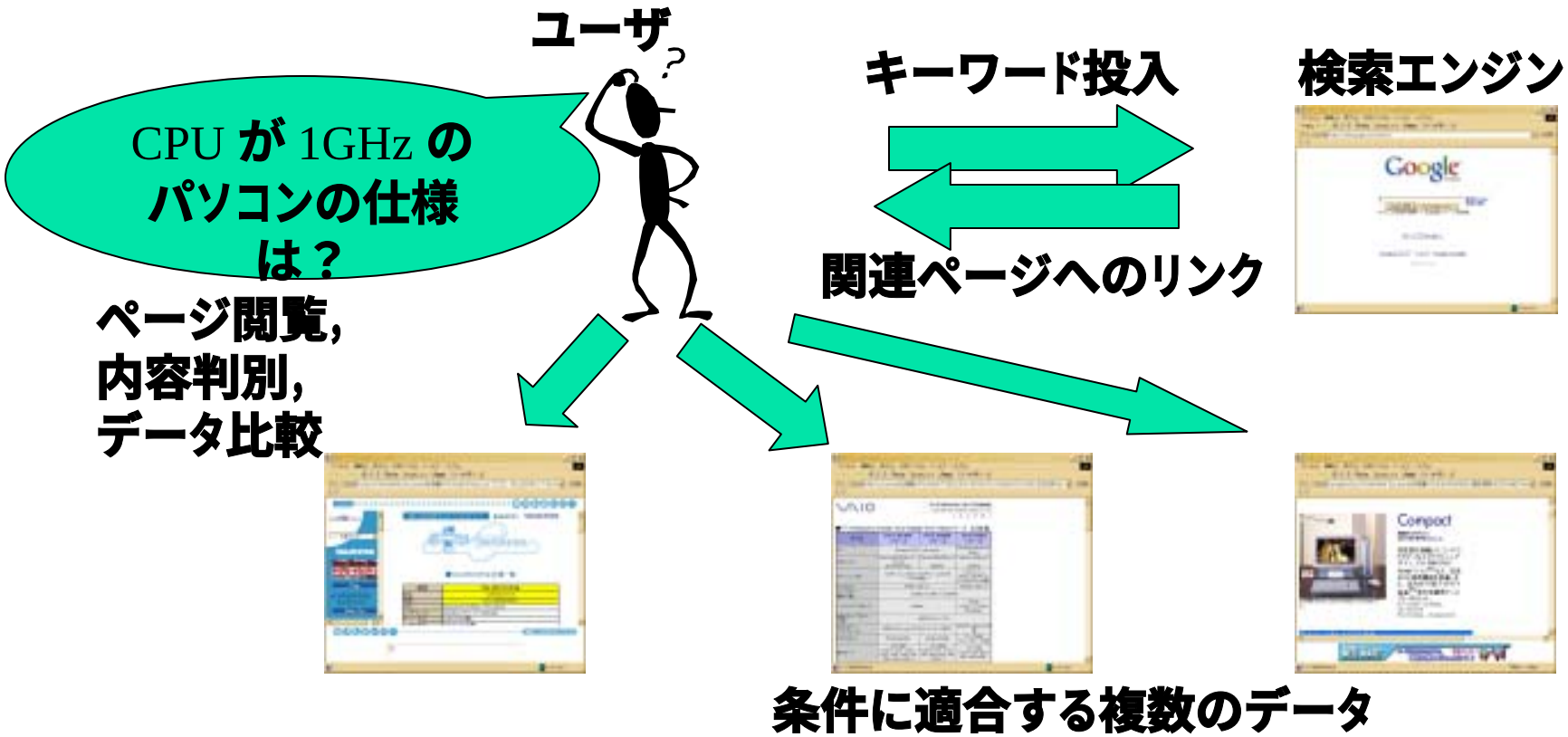
半構造データの差異発見

- そのままでは機械的な比較が困難
- 構造タグを利用 (OEM表現)
- 直感的に差異を理解できる表現が必要

Colored Bottom-up DataGuide



Web からの情報収集



データの比較･差異発見を自動化する

同属情報

同種の対象を同じ観点から特徴を整理したもの
(クラス)

同属情報の差異

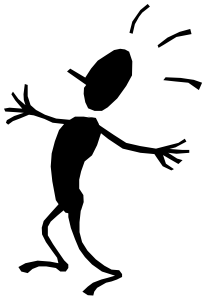
- 値の差異
  属性値の大小，一致／不一致
- 構造の差異
  属性の有無

(例)

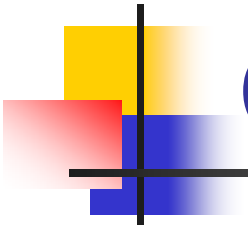
同属情報：パソコンの部品構成
値の差異：ＣＰＵ周波数の大小
構造の差異：ＤＶＤ-Ｒドライブの有無

提案する手法

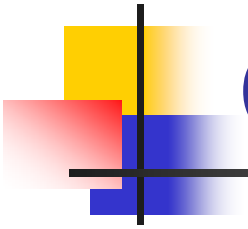
- Web 上のデータを OEM 形式で表現
- データの共通スキーマ生成
- データの特徴を可視化



OEM(Object Exchange Model)



- ツリー型のデータモデル
  - ノードに複数の親が存在してもよい
  - 循環する枝があってもよい
  - 枝はラベルがつく
  - ノードは識別子と値を持つ
    - 識別子によって一つ一つのノードは区別される
    - 中間ノードは子ノードを値として持つ
    - リーフノードは数値･文字列などの atomic な値を持つ



OEM への変形

- HTML タグを利用(Seung-Jin Lim らの方法)
  - 階層構造をツリー形式に変形
  - タグの意味も考慮
- 形態素解析を利用
  - 形態素解析によって文章から属性値を抽出
  - 属性値以外の文字列からラベルとして適当なものを選択 (属性値の単位など)

同属情報を OEM に変形

- 同属情報：
  パーソナルコンピュータの製品仕様（一部）

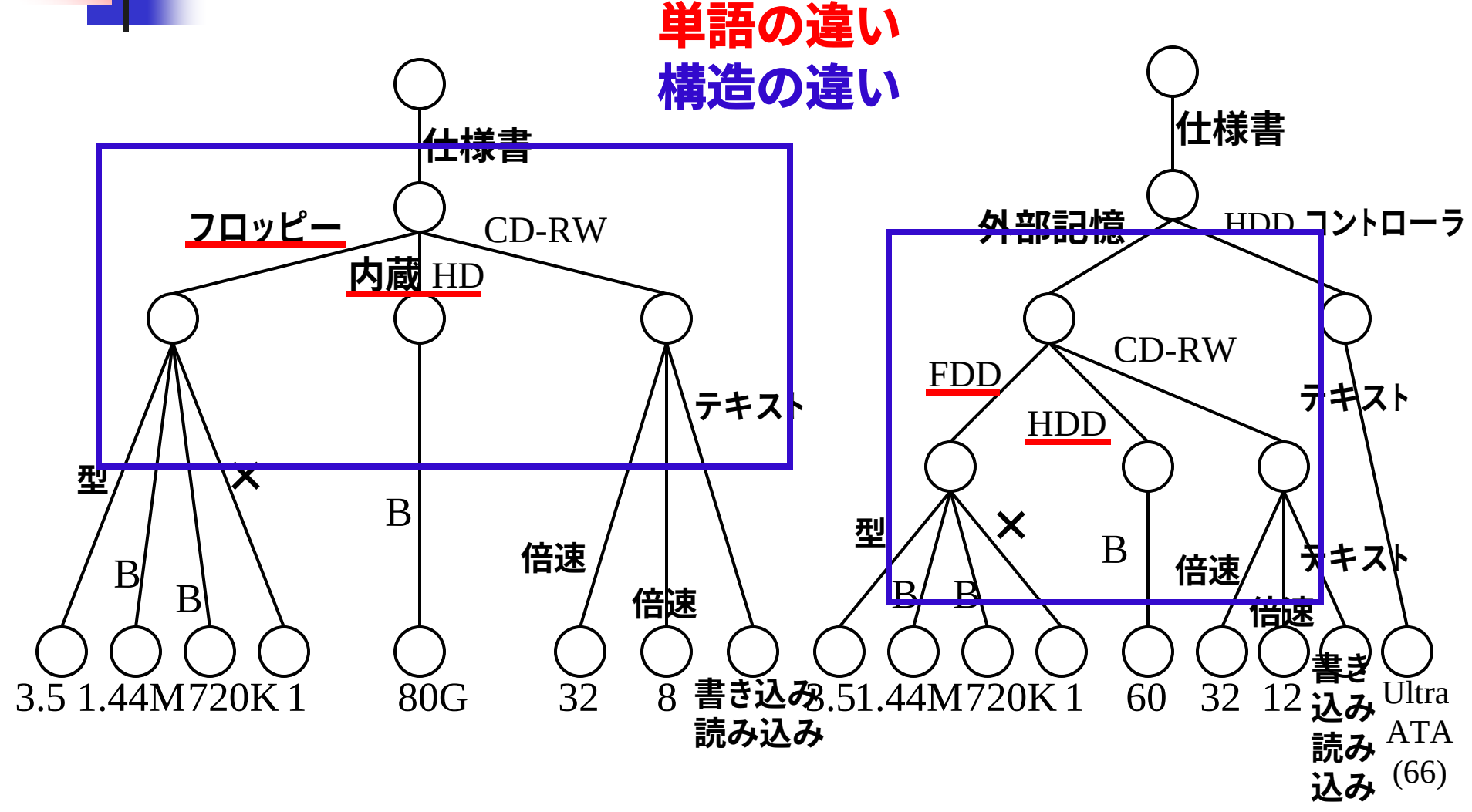
表 1：仕様書 1

項目	内容
フロッピー	3.5型 (1.44MB / 720KB) × 1
内蔵 HD	80 GB
CD-RW	読み出し 32 倍速 書き込み 8 倍速

表 2：仕様書 2

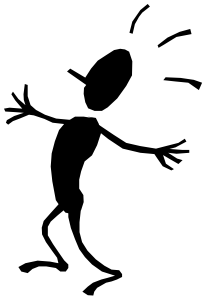
項目		内容
外部記憶	FDD	3.5 インチ (1.44MB / 720)KB ×1
	HDD	60GB
	CD-RW	読み出し 32 倍速 書き込み 12 倍速
HDD コントローラ		Ultra ATA (66)

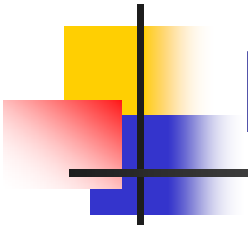
同属情報を OEM に変形



提案する手法

- Web 上のデータを OEM 形式で表現
- **データの共通スキーマ生成**
- データの特徴を可視化

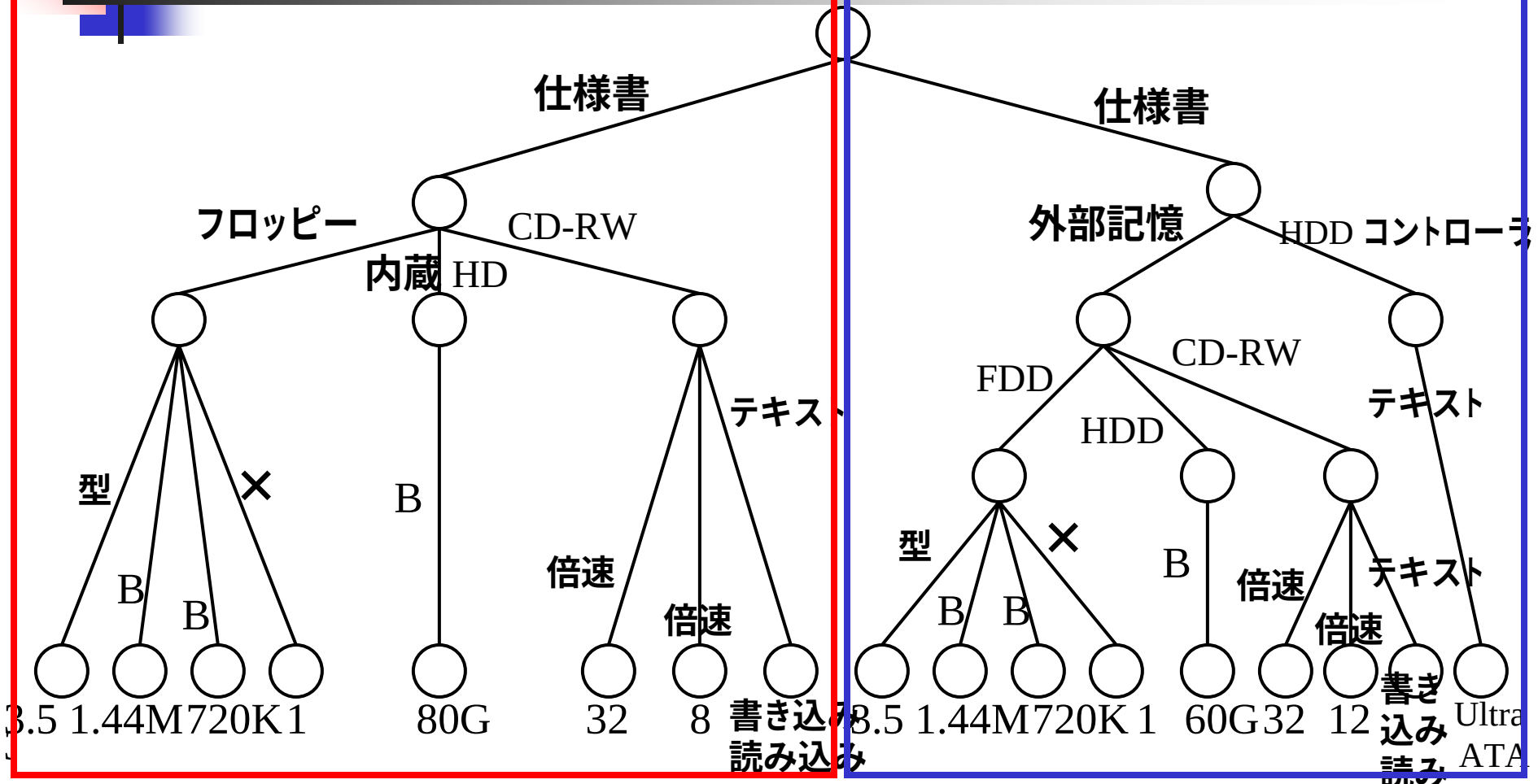




DataGuide

- DataGuide
  - OEM の同じ意味のノードを一つに集約して要約してできた OEM の総称
    - 到達可能な子ノードの集合 (ターゲットセット) によって共通のノードを判別
    - 集約したノードを DataGuide オブジェクトと呼ぶ
  - データごとにパスの異なる同じ属性値を指定可能
    - 集約した／されたノードの関係はハッシュ表に格納
- 近似的 DataGuide (Approximate DataGuide)
  - 類似なノードを一つに集約
  - 厳密な DataGuide より小さくなる
  - 類似の条件によって構造が変化

(近似的) DataGuide の問題点

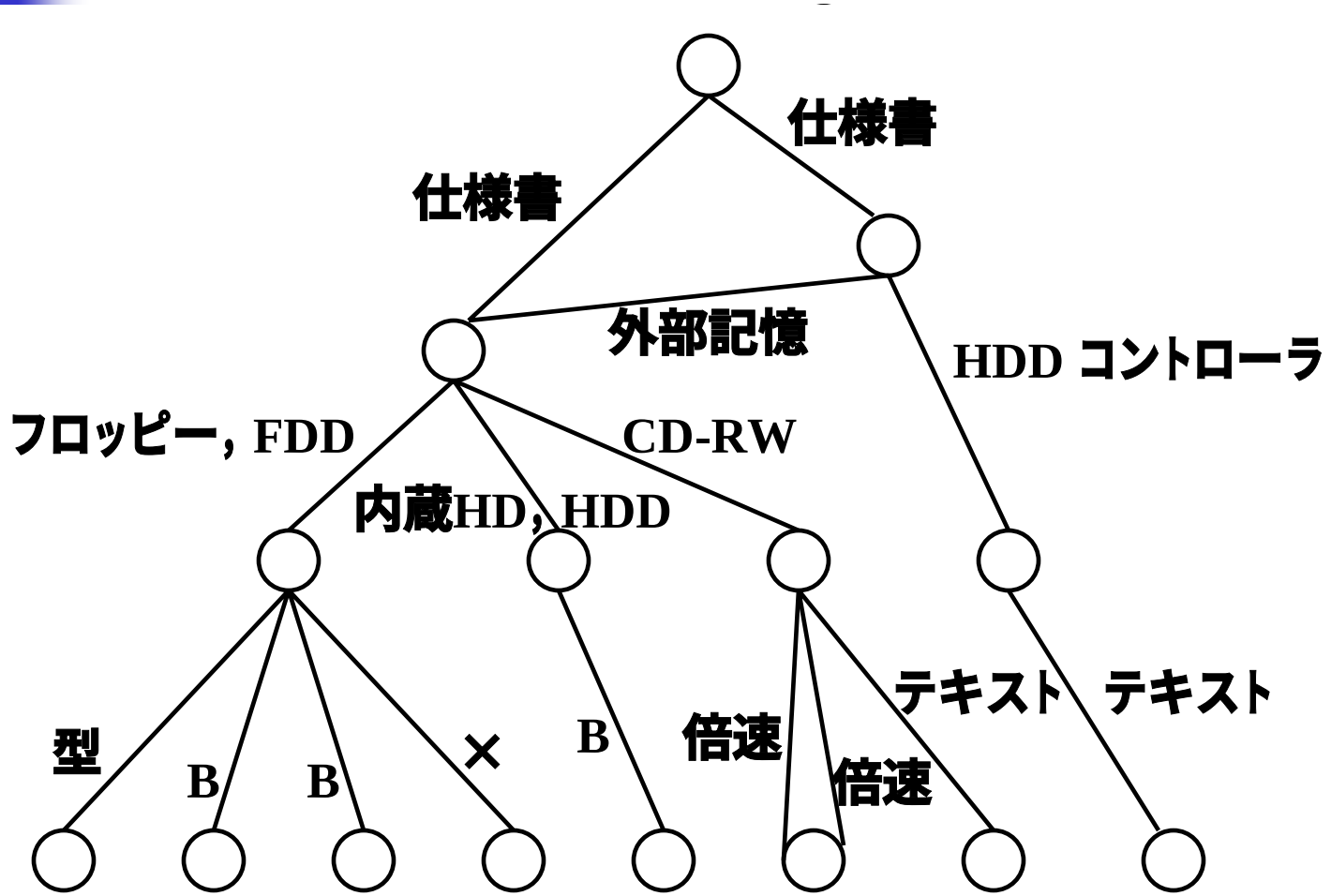


一致するターゲットセットを発見できない

BA-DataGuide (Bottom-up Approximate DataGuide)

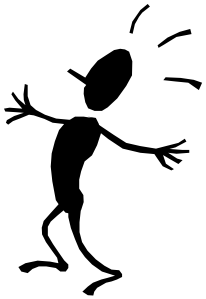
- リーフからルートへ DataGuide オブジェクトを生成
- 同じ性質のリーフノードを一つに集約
  - 値のレベルでは共通の属性は同じ性質を持つ
    - データ型
    - 単位
    - 値のとりうる範囲
- 子ノードが全てDataGuideオブジェクトで，ターゲットセットが類似なノードを一つに集約
- ノードを集約するときにメタデータを計算

BA-DataGuide の例



提案する手法

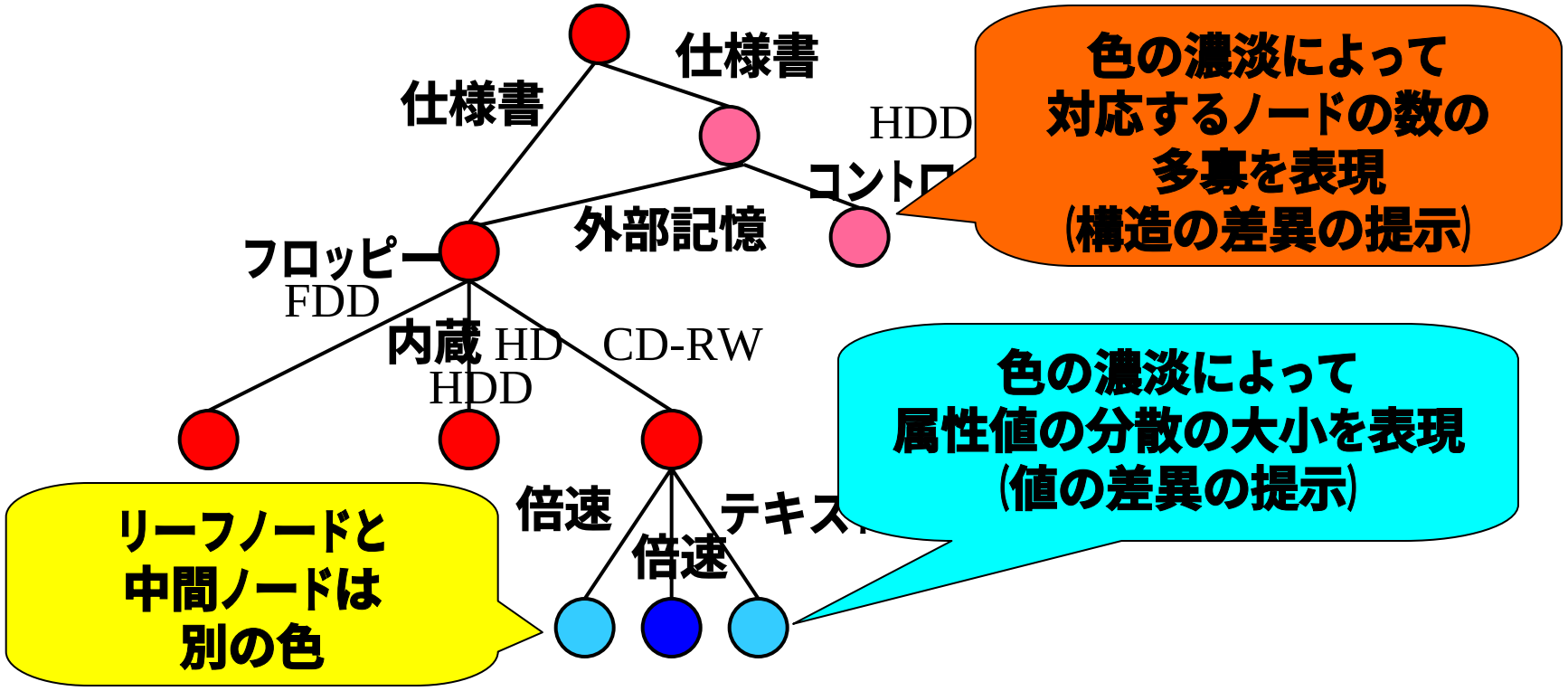
- Web 上のデータを OEM 形式で表現
- データの共通スキーマ生成
- **データの特徴を可視化**



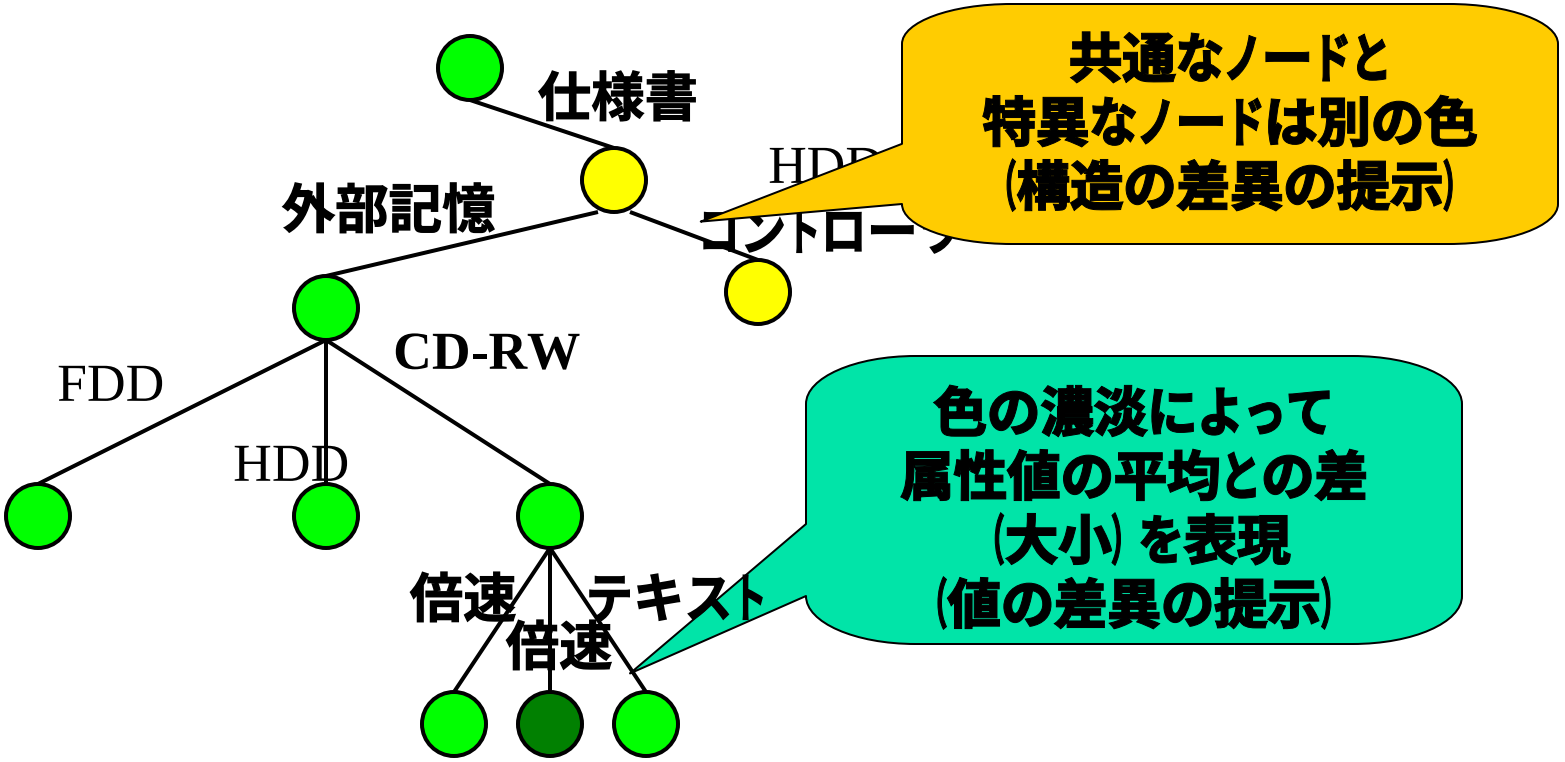
差異の可視化
(Colored Bottom-up Approximate DataGuide)

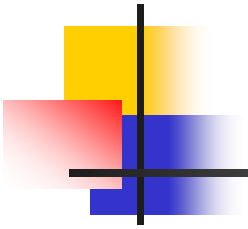
- ユーザにツリー形式でデータを提示
  - ツリーは自由に格納･展開可能
  - 共通スキーマの提示
    - データ全体の傾向を表現
  - 個々のデータのツリー表現を提示
    - 個々のデータの特徴を表現

差異の可視化 (収集データ全体)



差異の可視化 (個々のデータ)



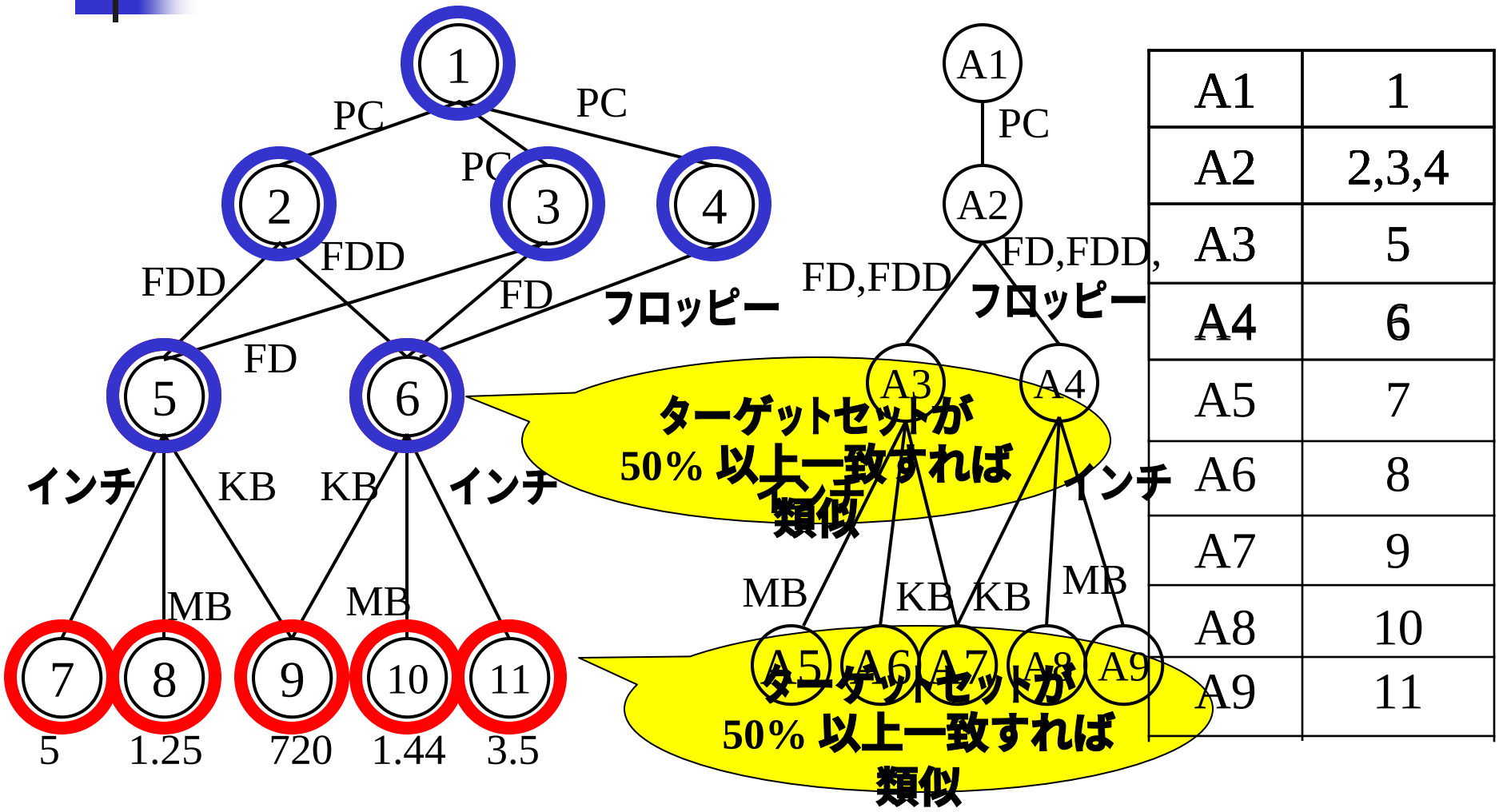


まとめ

- Web 上のデータを OEM に変形
  - データを共通の形式で表現
- 比較対象のデータの共通スキーマ生成
  - データごとの用語･構造の違いを解決
- 値･構造の差異の視覚化
  - データの共通･特異点の直感的な把握を支援

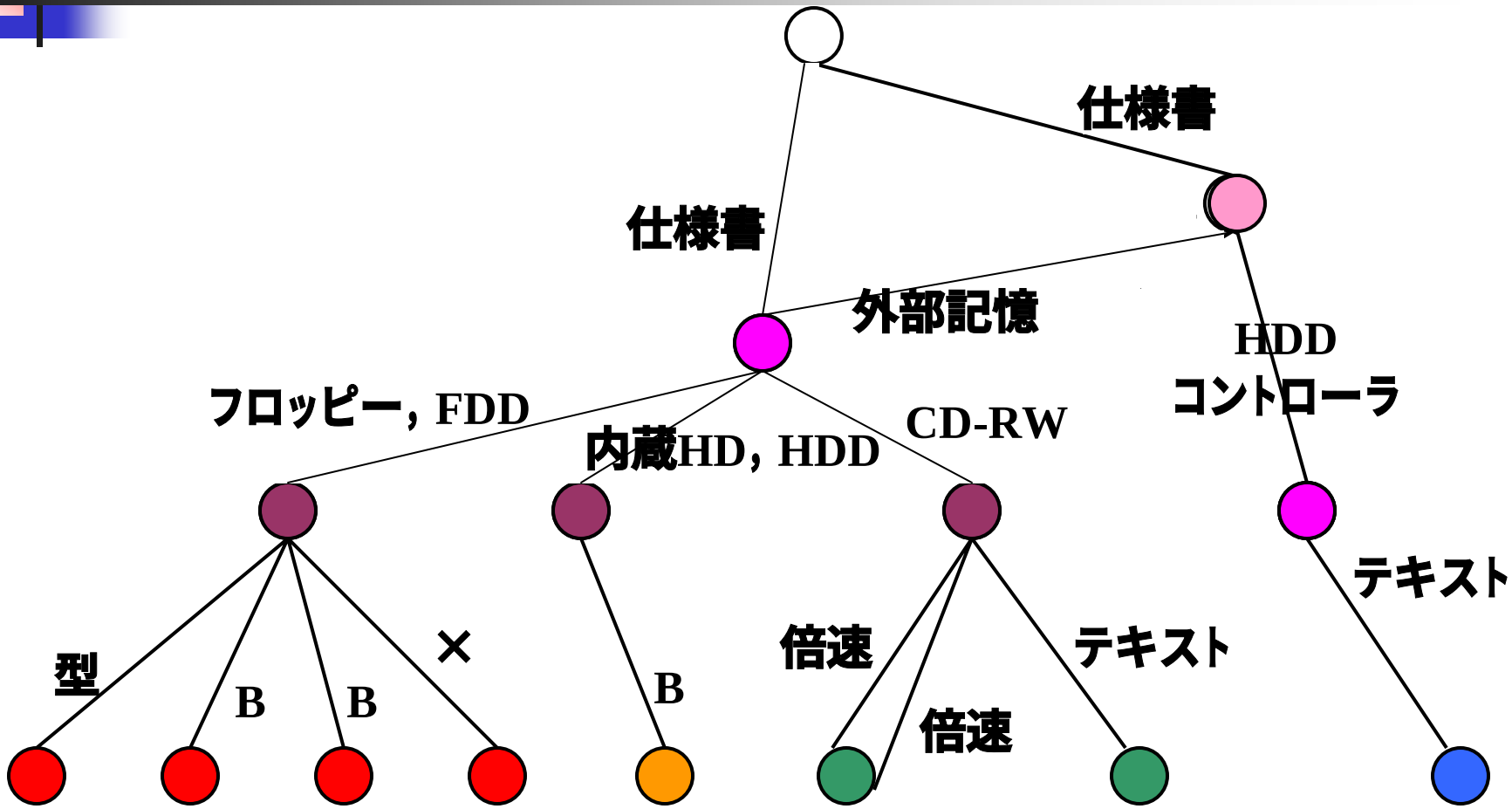
Web 上の半構造データの差異の発見･可視化

(近似的) DataGuide



A1	1
A2	2,3,4
A3	5
A4	6
A5	7
A6	8
A7	9
A8	10
A9	11

BA-DataGuide の例



関連研究

- SCD
  - 二つの Web ページ間の意味的な差異を検出
  - ページの更新前後の違いの発見に適している
  - 異なったページの比較には不適
- XWRAP
  - 複数のページに記載の情報の形式を統一するラッパー
  - ユーザとのインタラクションを繰り返してルールを生成
  - あらかじめデータ元のページから情報を抽出するルールを生成しておく必要がある