構造の異なる文書データを同じ構造に変換するアルゴリズムについて

岡山県立大学情報工学部
鈴木　伸崇

背景

半構造データの急速な増加

→データ構造が不規則で一定でない

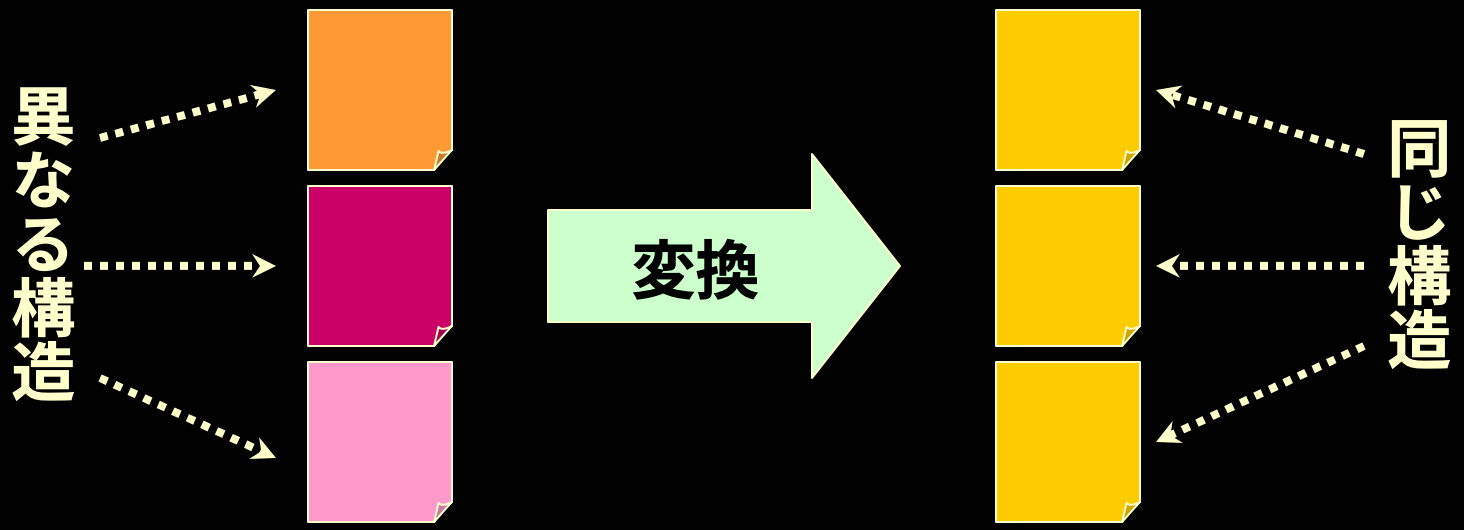
- ■データ格納・検索などの効率悪化
- ■質問記述の複雑化

データ構造は一定である方が望ましい

研究目的

下記データ変換を行うアルゴリズムの開発

構造の異なる文書データを，元のデータ構造を出来るだけ損なわずに同じ構造に変換



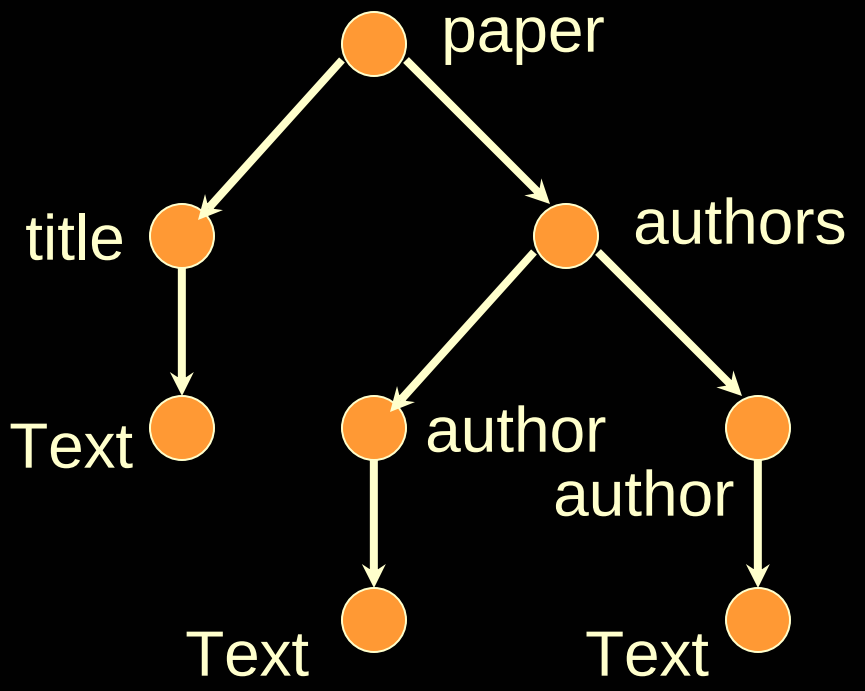
文書データ（1）

文書データを根付き順序木としてモデル化

- 要素が頂点
- 要素名が頂点のラベル

文書データ（2）

```
<paper>
   <title>
      Merging XML Documents
   </title>
   <authors>
      <author>
         Yoichirou Sato
      </author>
      <author>
         Michiyoshi Hayase
      </author>
   </authors>
<paper>
```



データ変換

以下の操作を用いて，構造の異なる文書データを同じ構造に変換（各操作にコストを与える）

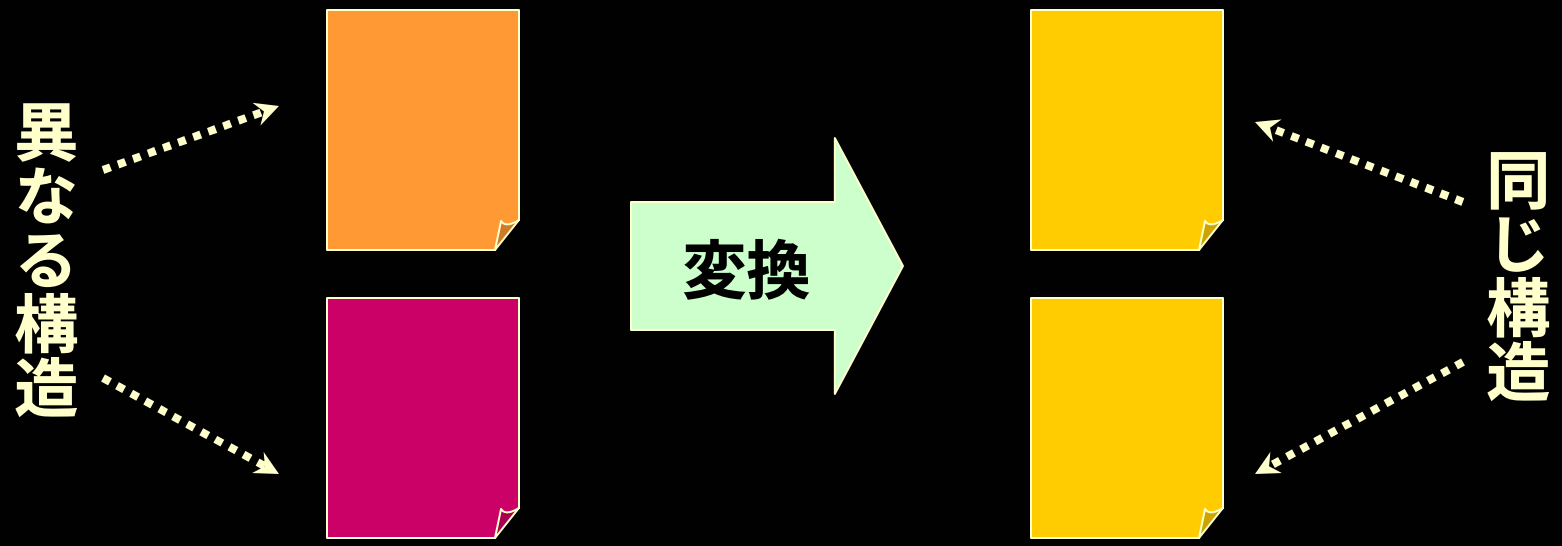
- 頂点の追加
- 頂点の削除
- ラベルの変更

変換コスト：文書データを同じ構造に変換するために必要な操作コストの和

変換コスト小　→　元のデータ構造がより保持される

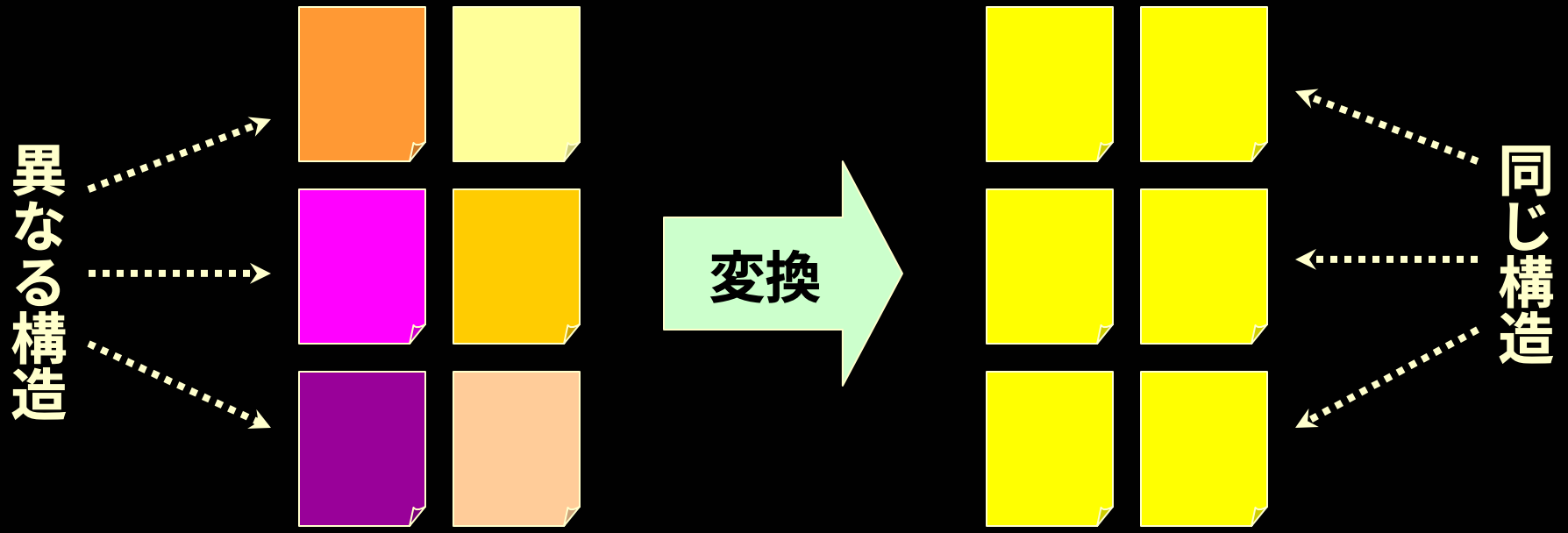
結果（1）

「２個の文書データを最小の変換コストで同じ構造に変換する」多項式時間アルゴリズムを開発



結果（2）

「任意の個数の文書データを最小の変換コストで同じ構造に変換する」アルゴリズムを開発



結果（3）

以下の問題がNP困難であることを証明

任意の個数の文書データを最小の変換コストで同じ構造に変換

多項式時間で効率良く解くのは困難

今後の課題

1. 以下の問題を解く多項式時間近似アルゴリズムの開発

    任意の個数の文書データを最小の変換コストで同じ構造に変換（NP困難問題)

2. 本アルゴリズム・上記近似アルゴリズムに関する評価実験

Thank you . . .