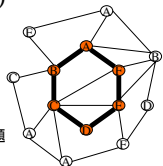
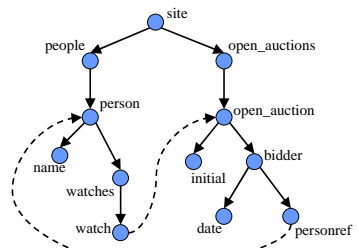


2. 研究の目的:

グラフと実データとの関係

グラフの利点 (と欠点)

- データのコンパクトな表現形式
- データの関係性を利用した高速アクセス
- 反面, **基本演算が困難**であることが多い(比較, 編集など)



部分グラフ同型問題

5

2. 研究の目的:

照合・検索・発見

■ 情報検索用語の再確認

- 照合**: データを前処理しないライトウエイトな手法
- 検索**: 補助情報(索引)を用いた高速応答
- 発見**: 正解が定まっていない曖昧な検索

■ コストと速度のトレードオフ

- 前処理をしない照合は省メモリ
- 前処理をする検索は高速応答
- 発見は時間・空間が高コストになりがち



イソップ寓話の挿絵 (wikipediaより転載)

6

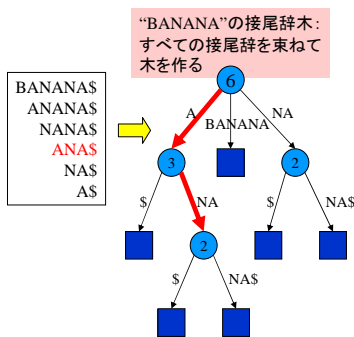
2. 研究の目的:なぜグラフ上の検索か?

グラフを利用した超高速検索の必要性

■ 従来の索引構造 (接尾辞木の例)

- 前処理をして索引を構築
- 一度索引ができれば以後は応答が高速
- バッファが $\Theta(n)$ 領域必要**

問題点: 大規模データの索引付けは困難

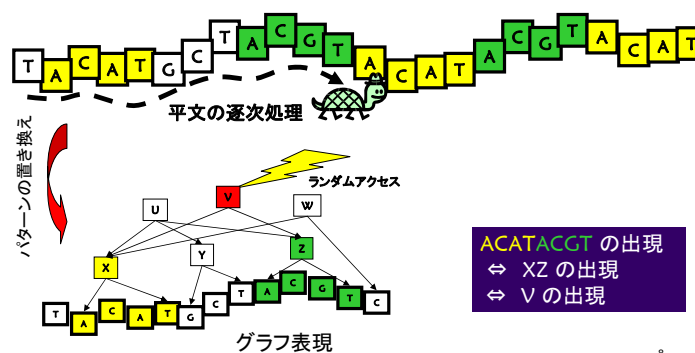


7

2. 研究の目的:なぜグラフ上の検索か?

省スペースデータベースの実現

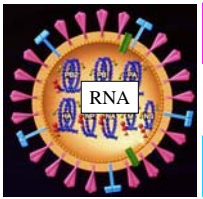
■ グラフ変換によるデータの圧縮



8

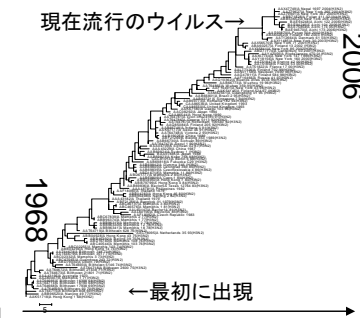
3. グラフ検索の研究事例 インフルエンザの抗原変異予測(1)

- インフルエンザウイルス
- H3N2亜型の進化系統樹



ヘマグルチニン (HA)
細胞への侵入の機能を司る

ノイラミニダーゼ (NA)
細胞からの遊離の機能を司る



現在流行のウイルス → 2006

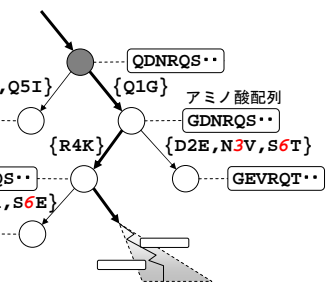
1968 ← 最初に出現

時間

提供: 伊藤公人先生
北海道大学人獣共通感染症リサーチセンター

3. グラフ検索の研究事例 インフルエンザの抗原変異予測(2)

- 抗原変異の仕組み
- パターンの抽出



QDNRQS...
Q1G
GDNRQS...
D2E, N3V, S6T
GEVRQT...
R4K
Q5I
S...
S6E

アミノ酸配列

(I) アミノ酸残基位置での共変異パターン
例 {D2G, Q5I}, {Q1G}, {R4K}, {D2E, N3V, S6T}, {N3R, S6E}


3番目の残基が変われば、6番目も換わる

(II) 系統樹の樹形パターン
例 {Q1?}, {R4K}

1番目がQ以外に変われば、4番目のRはKに換わる

3. グラフ検索の研究事例 インフルエンザの抗原変異予測(3)

- 進化の予測へ
- [Russell: *Science* '08] や [Igarashi: *Virology* '08] など



1968年 (香港かぜの出現) 2006年 (現在) 200X年 (近い将来)

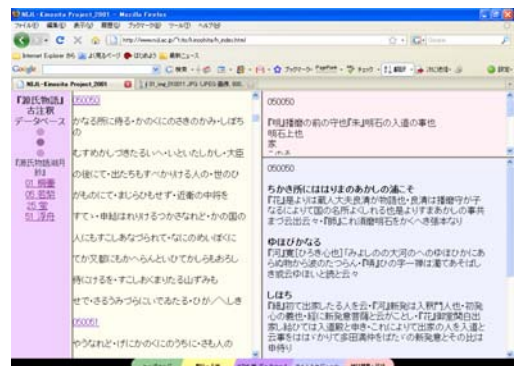
過去の抗原変異 将来の抗原変異

アミノ酸置換

共通パターンを発見 → 将来の変異を予測

3. グラフ検索の研究事例 異表記注釈付き古典文学の検索(1)

国文学研究資料館の源氏物語注釈データベース
<http://www.nijl.ac.jp/~t.ito/kinoshita/>



3. グラフ検索の研究事例
異表記注釈付き古典文学の検索(2)

古典的文書検索問題

入力: キーワード論理式 F とテキスト T
出力: テキスト T が F を真にするか否か?

$T =$ ぼんやりうさぎのさかもり

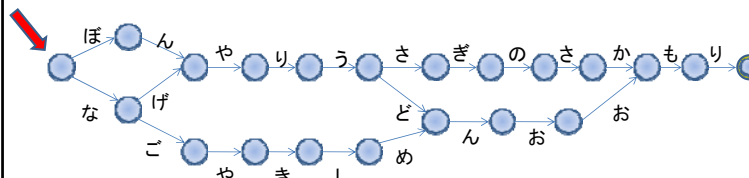
$F =$ ぼんやり and さかもり

13

3. グラフ検索の研究事例
異表記注釈付き古典文学の検索(3)

拡張文書検索問題はちょっと大変!

入力: キーワード論理式 F と有限な受理言語をもつ NFA A
出力: F を真にするテキスト T が $L(A)$ に含まれるか否か?

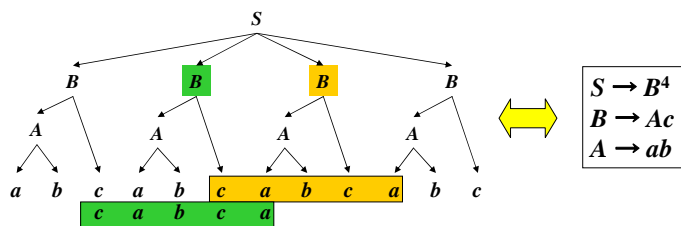


$F =$ ぼんやり and さかもり ○
 $F =$ ぼんやり and きしめん ×

提供: 竹田正幸先生
九州大学大学院システム情報科学研究院

3. グラフ検索の研究事例
巨大テキストの軽量索引

- パターンをうまく置き換える省メモリ圧縮
 - 理論的に最良の圧縮法 [Sakamoto: *IEICE J.*'09]



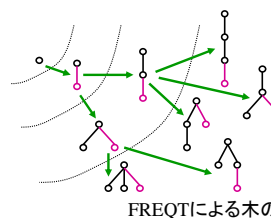
パターンの出現位置によらない
十分に長い部分文字列の置き換え

15

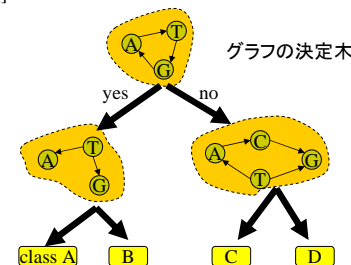
4. 関連研究:
グラフマイニング(1)

- 高頻度部分構造の抽出
 - 幅優先探索 (BFS)
 - AGM [Inokuchi: *PKDD*'00]
 - Path-join [Vanetik: *ICDM*'02]
 - 深さ優先探索 (DFS)
 - PrefixSpan [Pei: *ICDE*'01]
 - FREQT [Asai: *SDM*'02]

- グラフデータの分類
 - グラフの決定木 [Geamsakul: *PAKDD*'03]



FREQTによる木の列挙



16

5. これからの展望



- XMLの高度な利用
- Dagstuhl seminar in 2008
 - Structure-Based Compression of Complex Massive Data—
 - Query Evaluation on Compressed Documents
 - In-Memory XQuery/XPath Engine on Compressed Texts
 - SXSAQCT and XSAQCT: XML Queryable Compressors
 - The XQueC Project: Compressing and Querying XML
- パターン発見への拡張
 - 抗原変異予測への応用
 - より大きなデータへの適用...



XMLの圧縮と応用に興味がある人々 21

<http://drops.dagstuhl.de/portals/index.php?seminr=08261>