秘密分散データを用いた医療データ検索システム

1235067　中村 巴

February 10, 2021

ネットワーク信号処理研究室

被災地での診療の問題

東日本大震災では津波により病院に保存してたカルテが消失

⇩

患者が服用していた薬が分からない

⇨ 適切な診療が行えない

カルテを保全し，災害時に活用

活用時の問題

ただ保全すれば災害時に活用できるわけでは無い

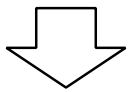
災害時には,

- ネットワークや電源などのリソース不足
- 必ずしもカルテの全てが必要とは限らない

カルテを保全し,
欲しい情報だけを入手できる仕組み

カルテを保全するための手法

カルテを保全するには，遠隔地に安全にバックアップ



- カルテ電子保存の 3 原則 (真正性，見読性，保存性)
- バックアップの条件 (秘匿性，冗長性)

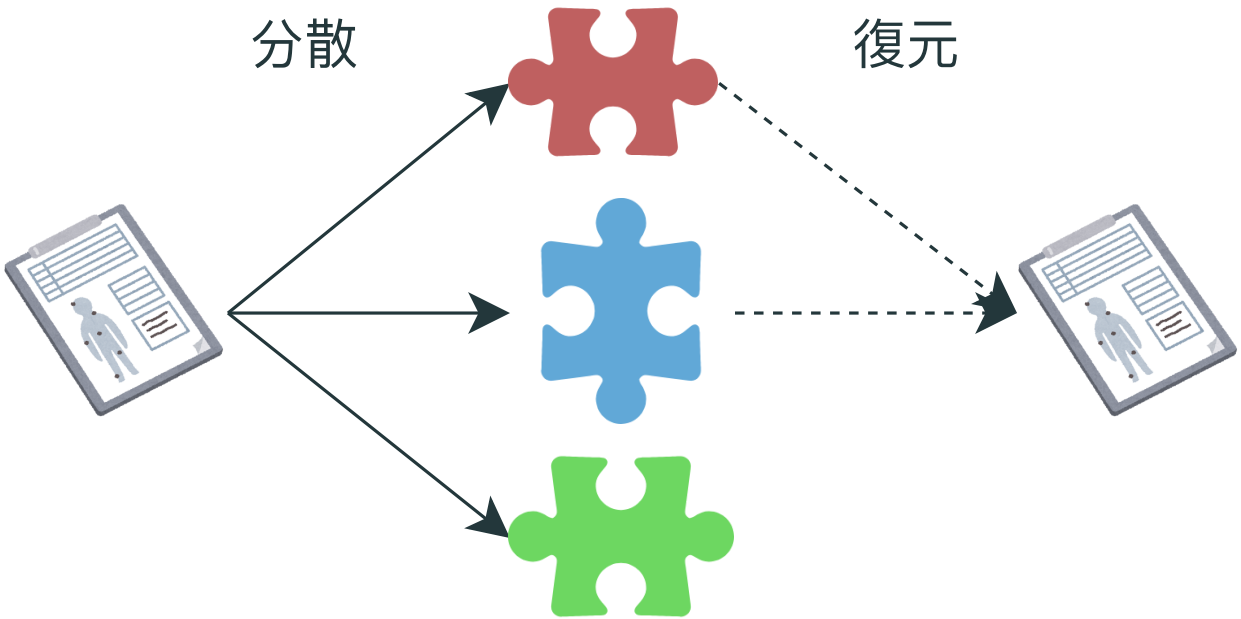
を満たす必要

(k, n) しきい値秘密分散法

(k, n) しきい値秘密分散法の問題

元のデータを復元するかしないかであり

欲しい情報だけを部分的に復元することはできない



部分的な復元が可能な秘密分散法の実現

カルテを保全し災害時に活用するための仕組み

部分復元可能な秘密分散法 [1]

秘密分散したデータの一部分だけを部分的に復元する.

- データを項目ごとに分割しそれぞれ秘密分散
- 欲しい項目のシェアを結合して復元

[1] 田中麻実，福冨英次，福本昌弘，“秘密分散バックアップした医療データの部分復元，” 信学技報 IA2015-74，pp. 31–36，2015.

部分復元可能な秘密分散バックアップした医療データを用いた診療

検索の仕組みは無い

⇩

名前などの識別情報のみを部分復元してから検索

部分復元可能な秘密分散法における検索の問題

- 医療データのシェアの組み合わせを知っている必要
  ⇨ 医療データ不正復元の可能性

- 復元は方程式を解くため $\mathcal{O}(l^3)$(医療データのデータ長を l) の計算量

- 全員分の識別情報を復元してから検索
  ⇨ 検索に時間を要するため，診療を始めることができない

目的と目標

目的

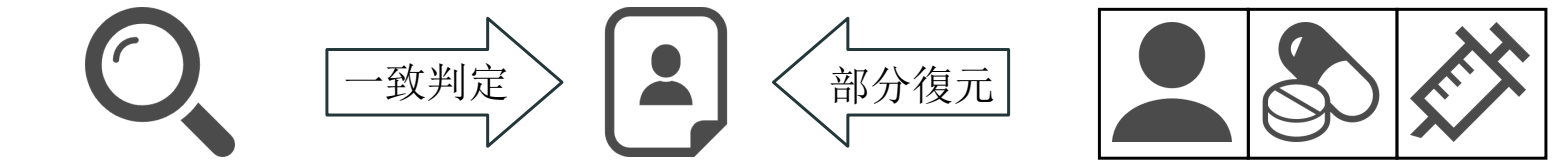
- 安全性を落とすことなく検索にかかる計算量の削減

目標

- シェアの状態での検索を実現
- 70 万データを 15 秒以内に検索

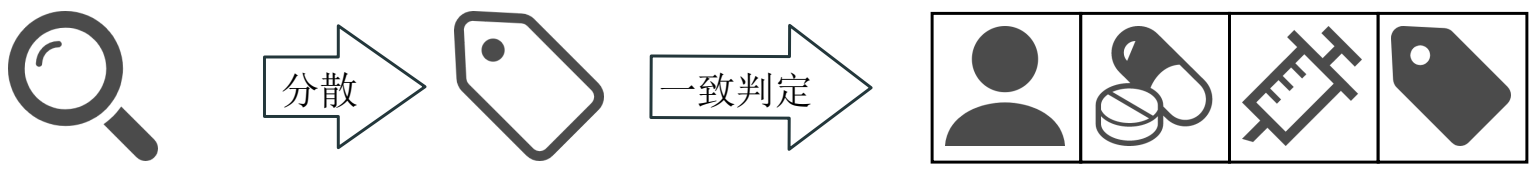
シェア間の係数の差の比較による検索

従来の検索方法



検索キーワード　　識別情報　　医療データのシェア

シェア間の係数の差の比較による検索



検索キーワード　　検索タグ　　医療データのシェア＋タグ

タグ，検索タグの作成方法

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} kw \\ r \end{pmatrix} \equiv \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix} \pmod{p}$$

保持

行列X 保持

タグ シェアに対応付けて保存

$$\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} kw' \\ r' \end{pmatrix} \equiv \begin{pmatrix} t'_1 \\ t'_2 \\ \vdots \\ t'_n \end{pmatrix} \pmod{p}$$

行列X

検索タグ

シェア間の係数の差の比較

タグ　　：$t_u \equiv kw + rx_u \pmod{p}$

検索タグ：$t'_u \equiv kw' + r'x_u \pmod{p}$

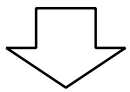
係数の差：$(t_u - t'_u)x_u^{-1} \equiv (kw - kw')x_u^{-1} + (r - r')x_u x_u^{-1} \pmod{p}$

$(u = 1, 2, \cdots, n)$

同じ添字のタグと検索タグを正しく比較

同一識別情報から作成したタグの組み合わせが分かる

⇨ 連立方程式の式の組み合わせが分かる



タグの組み合わせは分かるが，シェアの組み合わせは分からないような仕組み

シェアとタグの保存方法

Table 1: 添字が 1 のストレージ

医療データのシェア	タグ $_1$	タグ $_2$
A さんシェア $_1$	A さん t_1	A さん t_2
B さんシェア $_1$	B さん t_1	B さん t_2
⋮	⋮	⋮

各ストレージで一致判定ができ，シェアをランダムに保存することが可能

⇨ 一致判定に必要なタグの組み合わせは分かるが，2 個以上のシェアの組み合わせは分からない

検索時間の比較

- 識別情報，検索キーワードは 144bit(平仮名 9 文字)
- データ全体のサイズは 3000byte

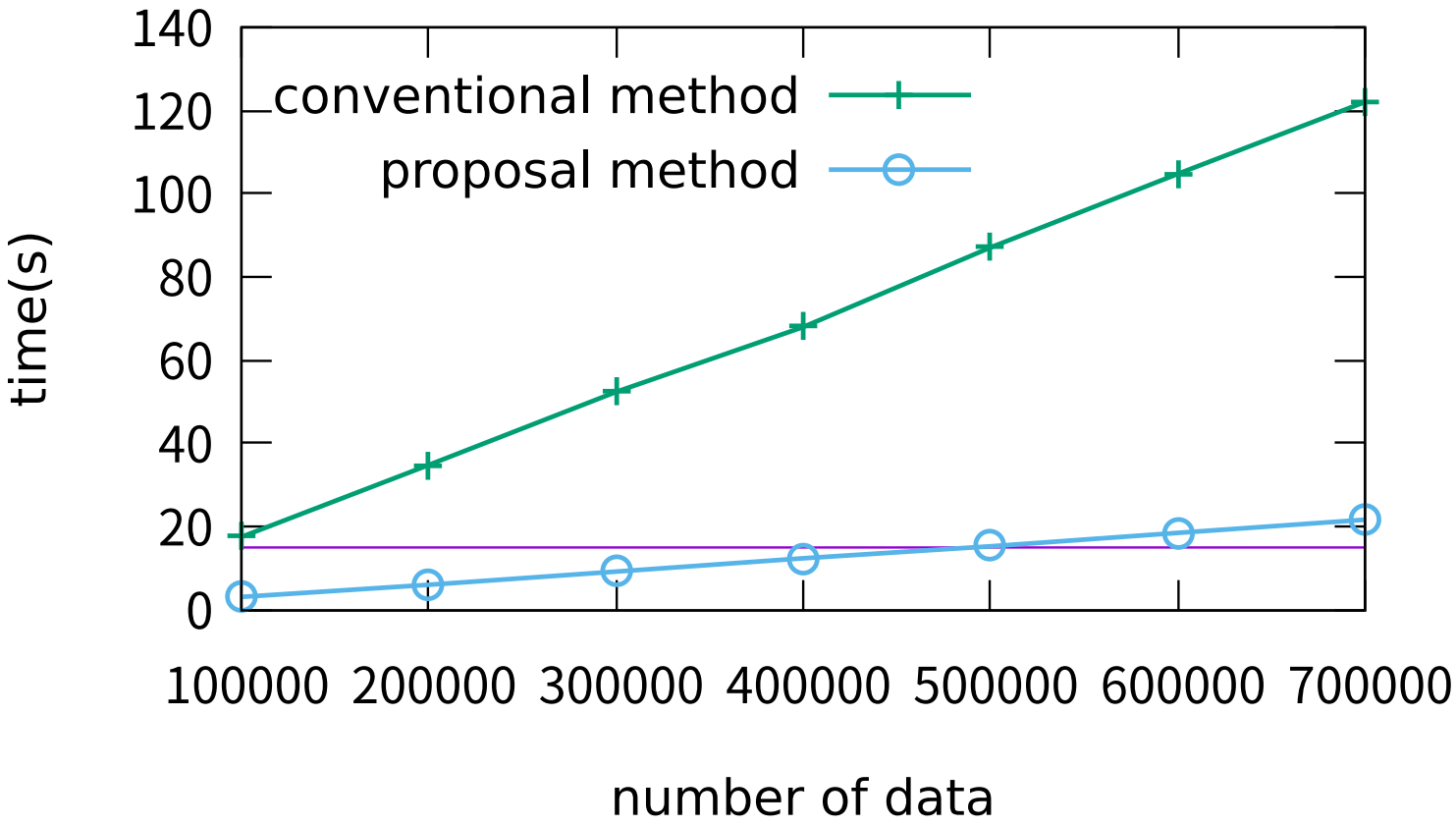
従来方法

識別情報を部分復元し，検索キーワードと一致判定

シェア間の係数差の比較による検索

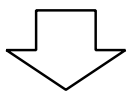
検索タグを作成し，タグと一致判定

結果


140
120
100
80
60
40
20
0
time(s)
conventional method
proposal method
100000
200000
300000
400000
500000
600000
700000
number of data

ここまでのまとめ

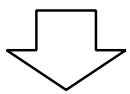
- シェア間の係数の差の比較による検索を提案
- シェアとタグの保存方法を提案



- シェアの状態で検索可能になり，計算量を $\mathcal{O}(d^2)$(識別情報のデータ長を d) に削減
- 検索時間は目標の 15 秒に大きく近づいた
- 第三者がストレージにアクセスしても医療データは復元不可

問題点

1. タグ t_1，t_2，しきい値 k，素数 p が漏洩
2. タグ t_1，t_2，しきい値 k，x_1，x_2 が漏洩



識別情報 kw が有限個に絞り込まれる

1. タグ t_1，t_2，しきい値 k，素数 p が漏洩

$$\begin{cases} t_1 \equiv kw + rx_1 \pmod{p} \\ t_2 \equiv kw + rx_2 \pmod{p} \end{cases}$$

解くには x_1，x_2 が必要

t の作成条件より x_1，$x_2 \in \mathbb{Z}/p\mathbb{Z} - \{0\}$，$x_1 \neq x_2$

$\Longrightarrow$ x_1，x_2 の組み合わせにより kw が絞られる

例

$$\begin{cases} 0 \equiv 2 + 3 \times 1 \pmod 5 \\ 1 \equiv 2 + 3 \times 3 \pmod 5 \end{cases}$$

t_1, t_2, p が漏洩↓

$$\begin{cases} 0 \equiv kw + r \times x_1 \pmod 5 \\ 1 \equiv kw + r \times x_2 \pmod 5 \end{cases}$$

kw について解くと，

$$kw \equiv \frac{0 \times x_2 - 1 \times x_1}{x_2 - x_1} \pmod 5$$

⇨ x_1，x_2 の組み合わせは 12 通り，$kw = 2$，3，4

評価

$$\begin{cases} t_1 \equiv kw + rx_1 \pmod{p} \\ t_2 \equiv kw + rx_2 \pmod{p} \end{cases}$$

kw について解くと，

$$kw \equiv \frac{t_1 x_2 - t_2 x_1}{x_2 - x_1} \pmod{p}$$

x_1，x_2 の組み合わせは $(p-1)(p-2)$ 通り

kw は $p-2$ 通り以下に絞られる

2. タグ t_1，t_2，しきい値 k，x_1, x_2 が漏洩

$$\begin{cases} t_1 \equiv kw + rx_1 \pmod{p} \\ t_2 \equiv kw + rx_2 \pmod{p} \end{cases}$$

kw について除算を行わず解き進め，p を総当たり

$\Longrightarrow$ ある条件下において，kw が有限個に絞られる

例

$$\begin{cases} 9 \equiv 7 + 2 \times 1 \pmod{11} \\ 0 \equiv 7 + 2 \times 2 \pmod{11} \end{cases}$$

t_1, t_2, x_1, x_2 が漏洩↓

$$\begin{cases} 9 \equiv kw + r \times 1 \pmod{p} \\ 0 \equiv kw + r \times 2 \pmod{p} \end{cases}$$

kw について除算を行わずに解き進めると，

$$kw \equiv \frac{9 \times 2 - 0 \times 1}{2 - 1} \equiv 18 \pmod{p}$$

$\Rightarrow (p = 11,\ kw = 7)\ (p = 13,\ kw = 5)\ (p = 17,\ kw = 1)$

$(p = 19,\ kw = 18)\ (p = 23,\ kw = 18) \cdots$

評価

$$\begin{cases} t_1 \equiv kw + rx_1 \pmod{p} \\ t_2 \equiv kw + rx_2 \pmod{p} \end{cases}$$

kw について解くと，

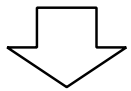
$$kw \equiv \frac{t_1 x_2 - t_2 x_1}{x_2 - x_1} \pmod{p}$$

$\dfrac{t_1 x_2 - t_2 x_1}{x_2 - x_1}$ がとる最大値は $(p-1)^2$

$\Longrightarrow$ kw は $(p-1)^2$ 未満の素数の個数通り以下に絞られる

考察

識別情報 kw が有限個に絞り込まれる

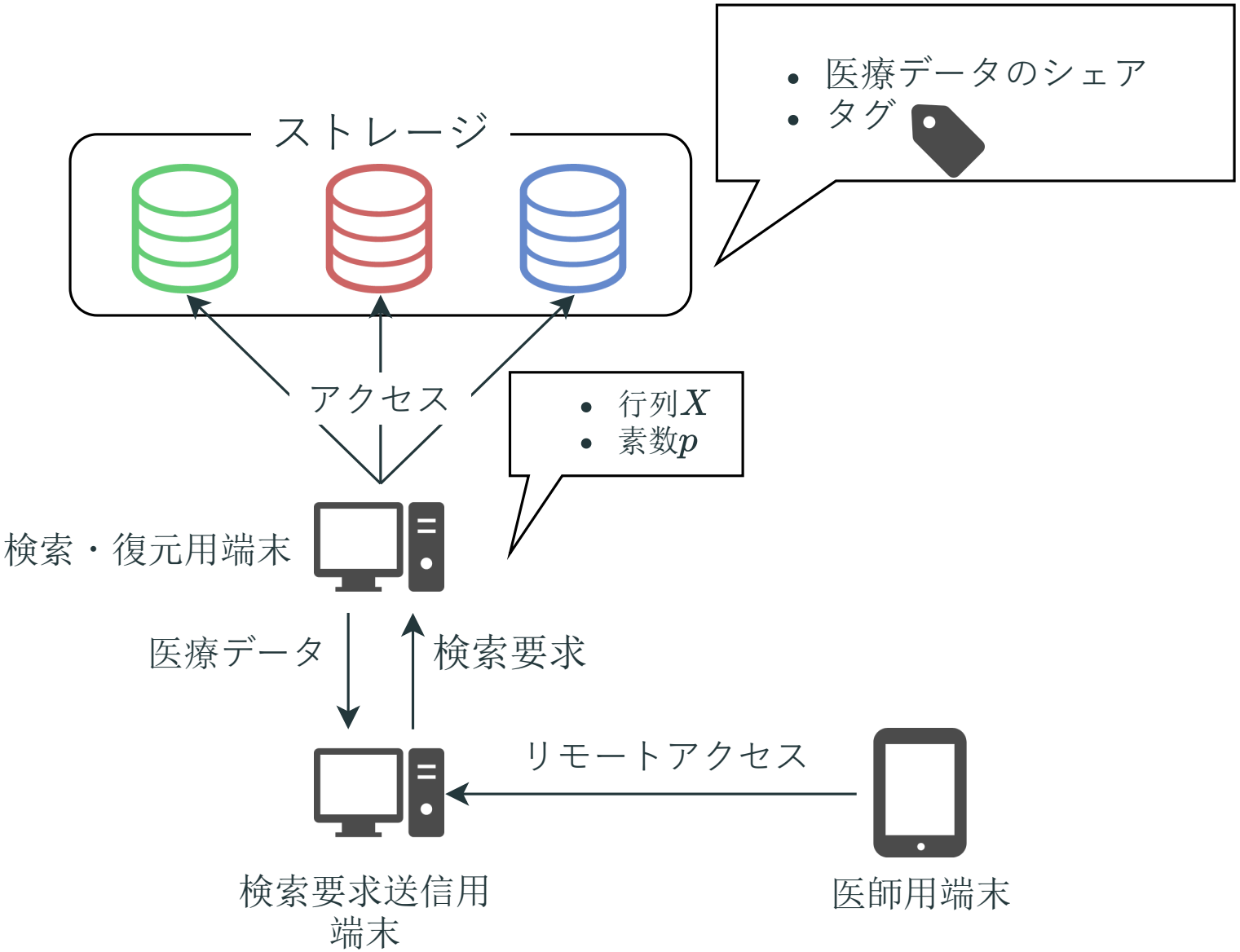


t, X, p をシステムで秘匿する

t
ストレージへのアクセス制限

X, p
拠点病院内のアクセスが管理できる場所に設置した端末に保存

t, X, p を秘匿したシステム構成



まとめ

- 目的
  - 安全性を落とさず検索にかかる計算量の削減

- シェア間の係数の差の比較による検索を提案
  - 計算量が $\mathcal{O}(l^3)$ から $\mathcal{O}(d^2)$ に削減
  - 検索時間は目標の 15 秒に大きく近づいた
- シェアとタグの保存方法を提案
  - 第三者がストレージにアクセスしても医療データは復元不可
- 安全性の評価
  - 識別情報 kw が有限個に絞り込まれる
  - システムで t, X, p を秘匿

- 今後の課題
  - さらなる高速化
  - システムを作成し，ネットワーク越しで実験