グローバルに情報をネットワーク化

柔軟な情報提供を可能にする次世代情報サービス基盤

慶應義塾大学 理工学部 情報工学科 金子研究室

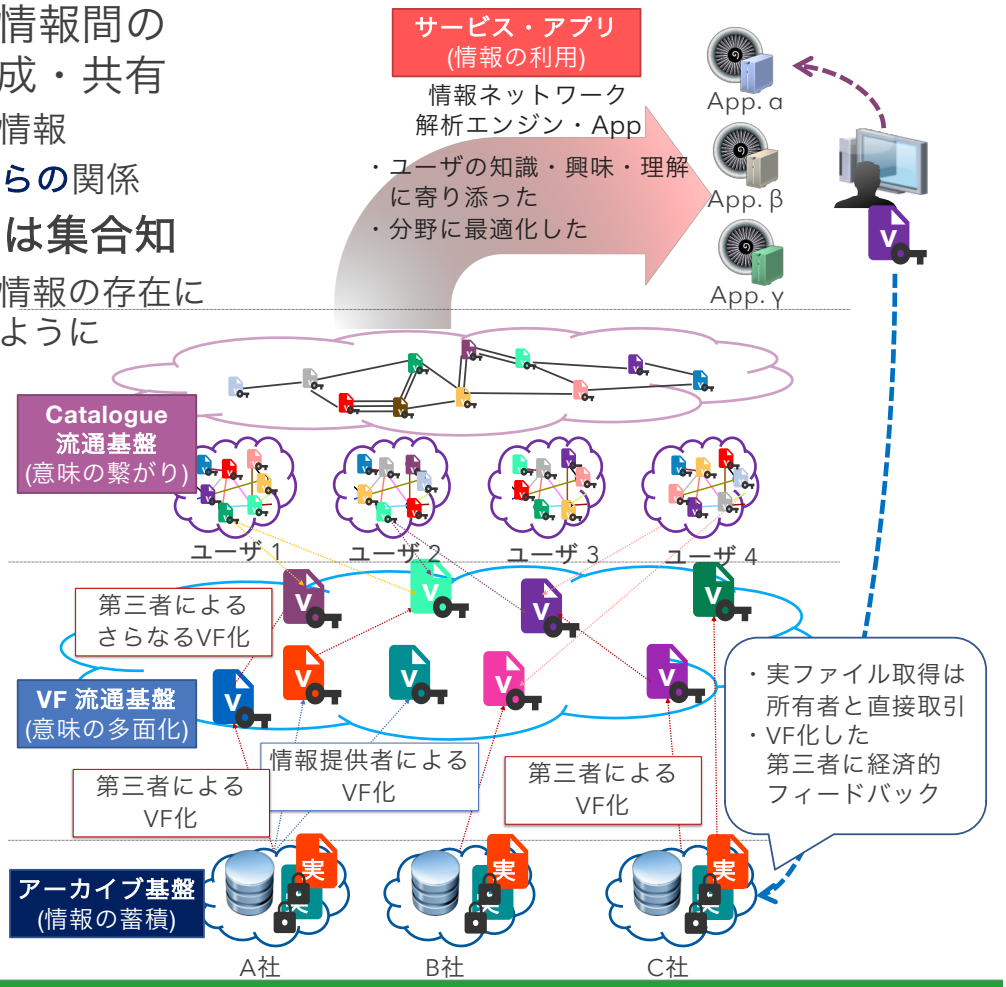
デジタル情報社会の今後

- デジタル情報の量の肥大化、数の増加
  - デジタル入力・出力デバイスの普及（例：センサのデジタル化）
  - デジタル加工に十分な計算資源（例：クラウド）
- 増加し続ける**デジタル情報を使いこなせるか？**
  - デジタル情報は、生成と保管に**コスト**がかかる資産
  - デジタル情報がなるべく**多く利用**されることが重要
- www+検索技術の限界
  - www+検索エンジンを基盤にしたより高度なサービスを構築できない

・キーワードは何？
・厖大なヒット件数
・情報反映に遅延あり
・特定事業者の寡占

ポスト検索：デジタル情報のネットワーク化

- **意味を規定したNW化ではなく意味の柔軟性を担保するNW化**
- 関係あるデジタル情報間のネットワークを構成・共有
  - ノード：デジタル情報
  - リンク：**なにかしらの**関係
- **情報ネットワークは集合知**
  - 関係あるデジタル情報の存在に速やかに気づけるように



研究トピック

- パーソナライズした情報提供のためのグラフ解析
- 規模拡張性(意味的・システム的)のある情報ネットワーク Catalogue System
- 保存と認証認可課金を流通から分離する Virtual File (VF)
- 情報蓄積の効率化

Catalogue System：
コンテンツ空間の相互接続

慶應義塾大学 理工学部 情報工学科 金子研究室

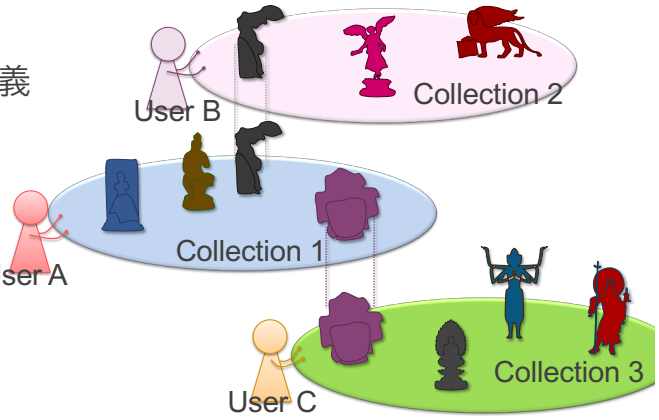
メタデータによるコンテンツ発見の課題

- メタデータ入力者と検索者の意味のずれ（メタデータ**語彙の共通化**が困難）
  - オントロジー導入が意味の混交を生み、検索精度の劣化を招くことも
- 無数のメタデータフォーマット（**フォーマット統一**は組み合わせ爆発のため困難）
- 串刺し検索組織の不在（**串刺し検索組織**にメリットなし）
- 串刺し検索組織へのデータ登録**共通ルール**の欠如
- ボーダレスなデジタルデータ流通に**統制型アプローチは不適合**
  - コンテンツ空間は自律分散的存在

Catalogue System

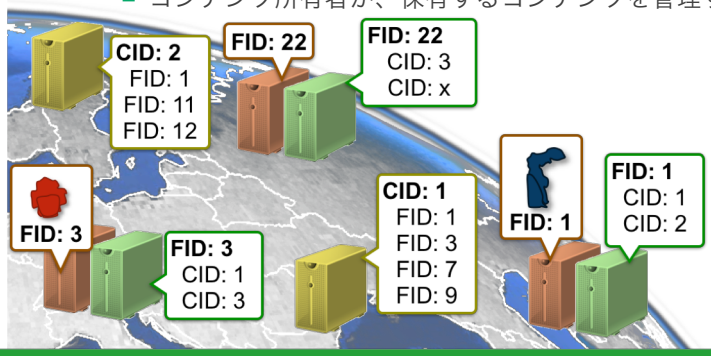
- メタデータを介した**間接的**なコンテンツ同士の結びつけではなく
  利用者の意思により**直接的**にコンテンツ同士を結びつける

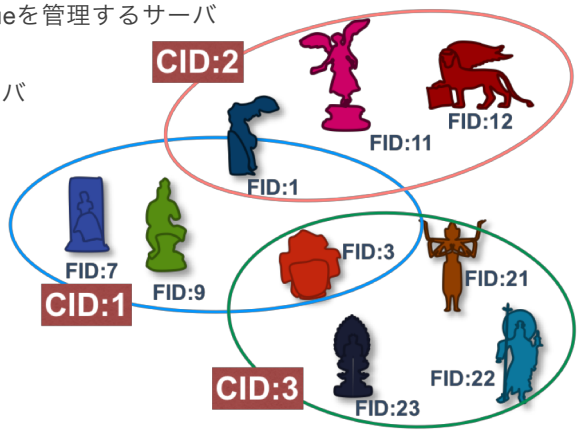
- メタデータを付与する代わりに、
  利用者が関係のあるコンテンツ集合を自由定義
  - 所有権のないコンテンツも結びつけ可能
- コンテンツがコンテンツ集合間を直接橋渡し
  - 自然言語の意味精度は、
    コンピュータの検索精度より圧倒的に低い
  - シンプルで汎用的な結びつけ機構
  - デジタルオブジェクトのID空間を揃える



システムの構成

- Catalogue Server (CS)
  - カタログ所有者が、保有するCatalogueを管理するサーバ
- Graph Manager (GM)
  - コンテンツ所有者が、保有するコンテンツを含むCatalogueを管理するサーバ
- File Server (FS)
  - コンテンツ所有者が、保有するコンテンツを管理するサーバ

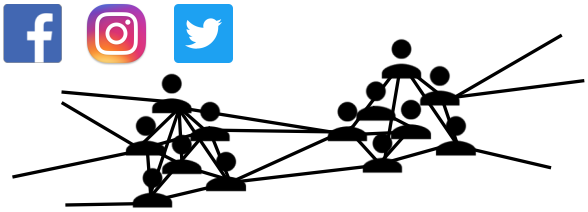




Catalogue Systemを活用した推薦システムの開発に向けて

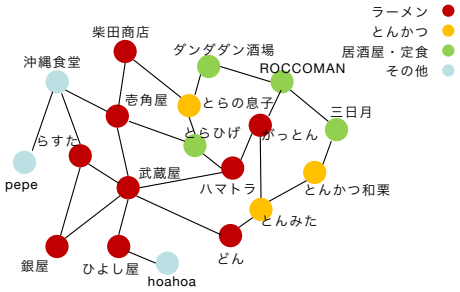
慶應義塾大学　吉谷 遼,志賀野 泰岳,米川 和亨,岡 亮, 金子 晋丈　{paul, pika, bass, okapy, kaneko}@inl.ics.keio.ac.jp

現在のグラフデータ活用のありかた



近年，1つ1つのデータだけでなく
データ間の繋がりを利用したデータ活用が増加
- e.g., SNSの友だち関係の解析・利用

Catalogue Systemでつくる新たなグラフデータ活用



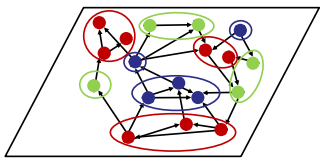
Catalogue Systemは個人個人が
それぞれの価値観で作成したグラフの集合

多様な意味を内在するグラフ情報を最大限活用した
アプリケーションとして日吉近辺の飲食店を題材に
推薦システムを研究・開発

自律分散グラフ探索の効率化

自律分散グラフの探索は
集中型グラフに比べ困難
- ノードが複数サーバに存在（分散）
- ノードの配置が最適でない（自律）

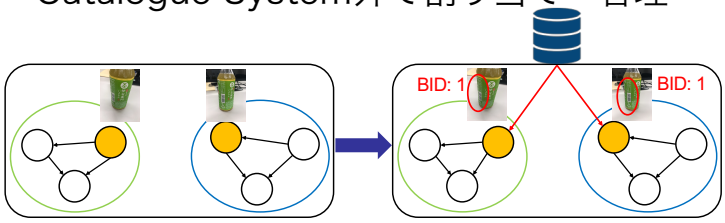
重要なノードを優先的に探索するなど
グラフ探索を効率化する
手法を研究中



同一対象を表すデータの紐付け

現状のCatalogueでは同一対象を表す
データが別のデータとして扱われている
→同一対象のデータが紐付いていない

データの同一性を表す識別子（BID）を
Catalogue System外で割り当て・管理

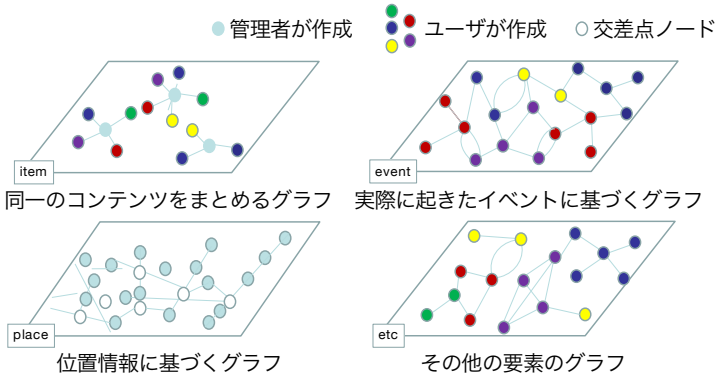


同一対象を表すデータが
紐付いていない

同一対象を表すデータが
紐付いている

グラフ特徴に基づく柔軟な探索手法

グラフを特徴ごとに分類



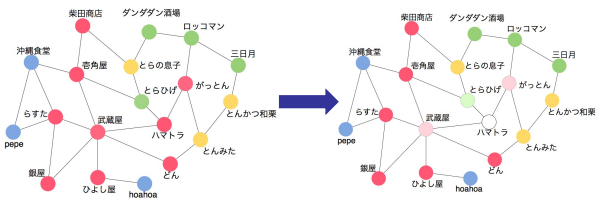
同一のコンテンツをまとめるグラフ
実際に起きたイベントに基づくグラフ
位置情報に基づくグラフ
その他の要素のグラフ

グラフが持つ特徴を並列制御し活用

フィルタリングによる推薦の効率化

除外したいコンテンツを指定したときに
類似コンテンツも除外

候補数を減らして推薦精度と効率を向上



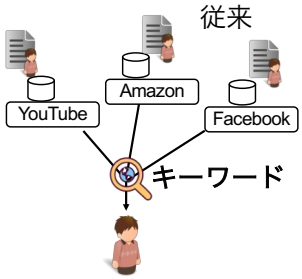
ハマトラを入力したときに武蔵屋も除外

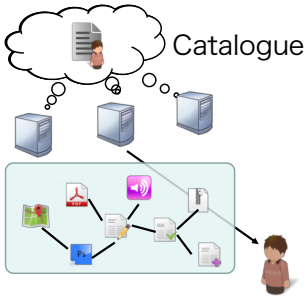
ユーザが作成した Catalogue に基づいたコンテンツ提示

慶應義塾大学 佐野岳史, 上村優介, 石川裕也, 青木佳奈, 金子晋丈 {luso, simba, sayuri, annie, kaneko}@inl.ics.keio.ac.jp

Catalogue の提示方法

- 従来のコンテンツ提示
  - キーワードで対象コンテンツを絞り人気コンテンツを提示
  - 各サービスドメインで集められたユーザの閲覧履歴からパーソナライズ化
- Catalogue を用いたコンテンツ提示
  - コンテンツの関係性で絞り込む
    - 関係のあるコンテンツを網羅的に提示
  - ユーザの所有 Catalogue によるパーソナライズ化

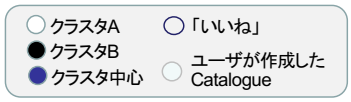




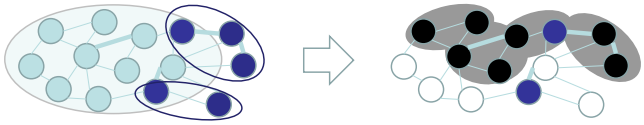
ユーザの知識や活用方法に合わせたコンテンツ提示を目指す

ユーザの興味指向性を考慮したCatalogue境界の発見

- 従来のクラスタリングではユーザの興味が反映されない
  - クラスタの密度をもとに画一的な境界を提示
  - ユーザ情報がサービス提供者ごとに分離
- 興味の中心からの意味的距離をもとにCatalogueを切り分けることでユーザの興味を反映
  - 興味中心：ユーザの知識と興味の境界
    - ユーザの自作Catalogue：ユーザの知識
    - 「いいね」したCatalogue：ユーザの興味
  - 意味的距離：次数の差とエッジ重複を考慮



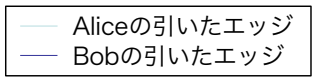
どのユーザに対しても同じ境界になる

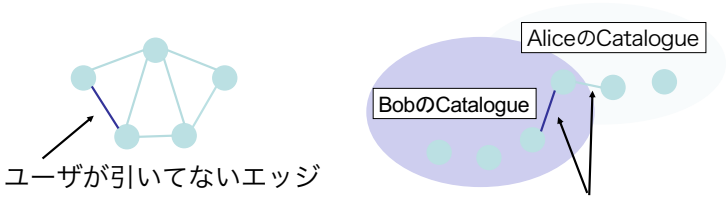


ユーザごとに興味の中心ベースで異なった境界を提示

意外なつながりの発見

- ユーザの知識をもとにユーザごとの意外なつながりを発見
  - ユーザが引いてないエッジ
  - ユーザが作成したCatalogue内のコンテンツとつながるエッジ
- 知らなかったコンテンツの側面を知り，知見を深めることが可能

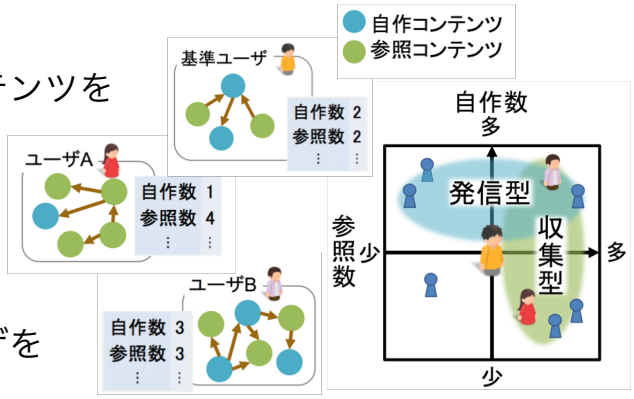




Catalogue内のコンテンツとつながるエッジ

相対的なユーザ類型判定

- 判定対象のユーザを基準とし，共通のコンテンツをCatalog化しているユーザと比較して，相対的にユーザの類型を判定
  - 比較指標：自作数，参照数，次数，被参照数 etc.



- 基準ユーザに対し，発信型，収集型のユーザを発見して，コンテンツ提示に利用

Virtual File

データの流通・共有・活用のためのプロキシフォーマット

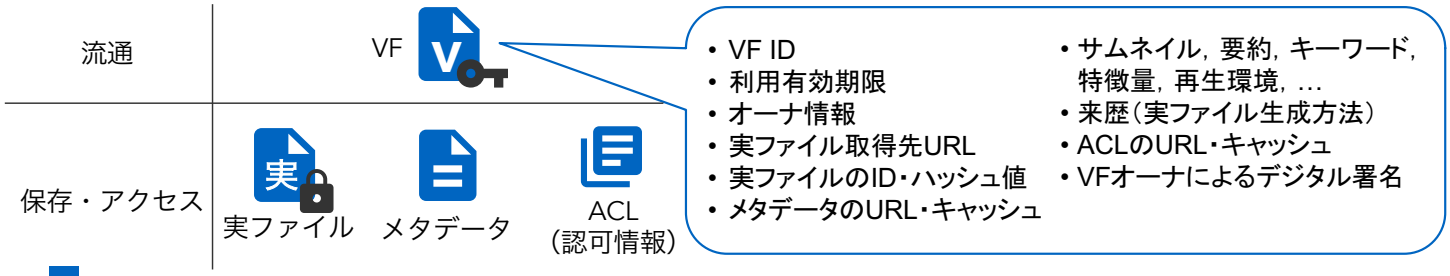
慶應義塾大学 荻谷 凌，金子 晋丈 {haru, kaneko}@inl.ics.keio.ac.jp

デジタルデータ利活用促進の課題

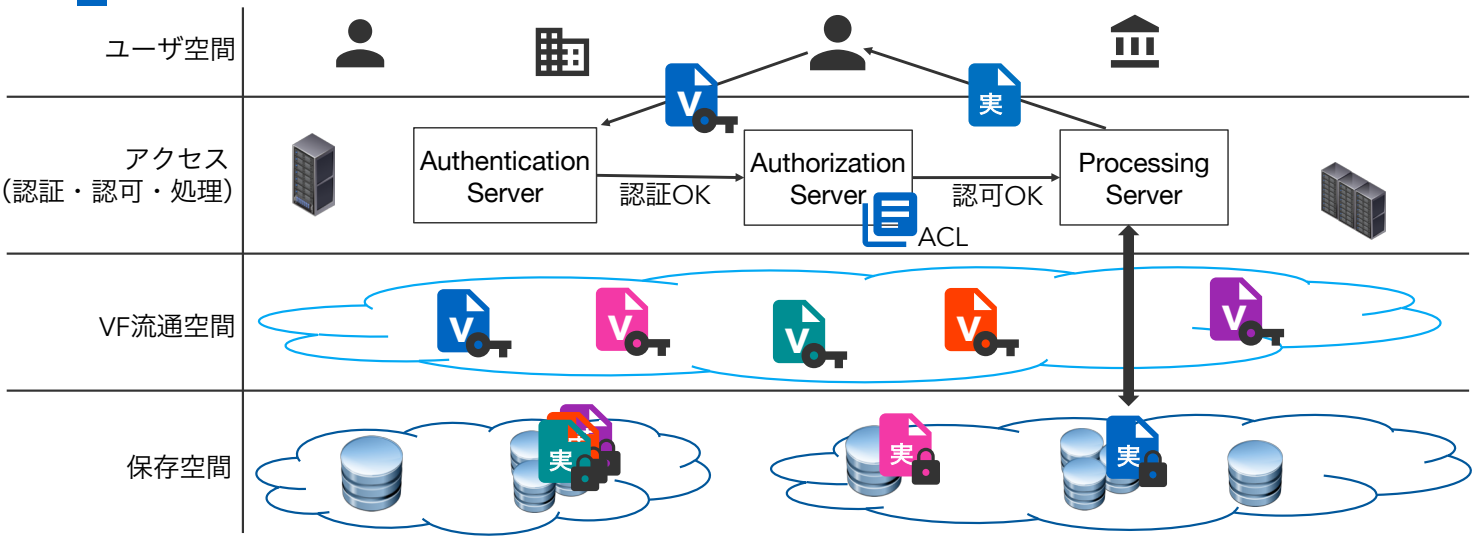
- 保存・流通・アクセスが分離されていない
  - **ファイルID空間**（URL・ファイル名・DB）の異なるデータの活用
  - **大容量データ**の流通
  - **多様なアクセス方式**（e.g., HTTP）のサポート
- 利用方法が不明
  - 統合利用する**複数データ**のパッケージ化
  - データの領域指定，**処理方法の共有・再現**
- 閲覧・処理権限が不明
  - **認可主体や利用条件**が不明

Virtual File (VF)

- **実ファイル**の**プロキシ（代理）**となる**仮想ファイル**（XML）
  - 軽量で，実ファイルの代わりに流通
  - ファイルの各種メタ情報を記載（e.g., ID，タイトル，サムネイル，キーワード，etc.）
- **Access Control List (ACL)** により，各ユーザの利用権限を制御
- 複数データを組み合わせた処理の場合，処理手順をVFに記載
- ハッシュ値・処理手順により，実ファイル・データ活用の再現性を保証



VFを用いたデータ利活用プラットフォーム



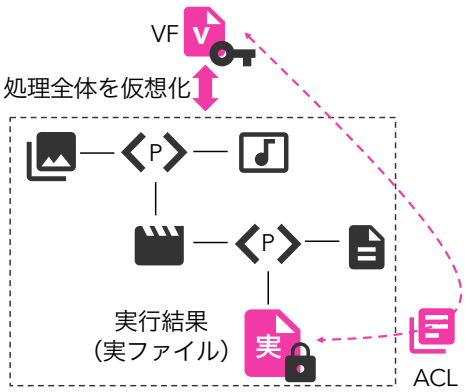
Virtual Fileの応用

データの保存・流通・活用のためのプロキシフォーマット

慶應義塾大学　荻谷 凌，趙 元韜，田部 悠介，金子 晋丈　{haru, arthur, al, kaneko}@inl.ics.keio.ac.jp

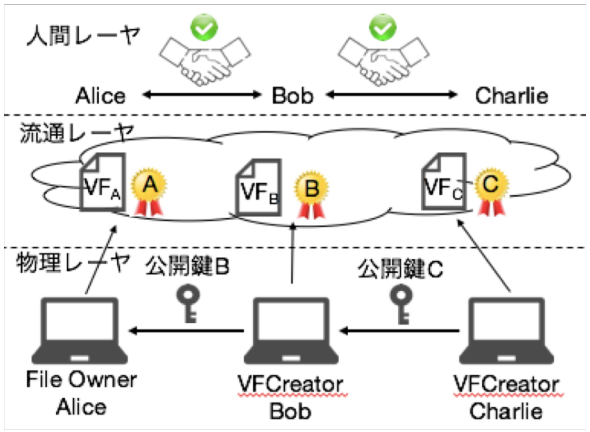
複数データの統合利用

- データ処理を含めて仮想化
  - データ処理方法の共有・再利用
  - データ保存量の削減
- 実データファイルの生成方法を来歴としてVFに記載
  - 環境（e.g., OS，依存アプリケーション）
  - 依存VF IDリスト
  - 実データファイル生成コマンド
- VF利用時，来歴に従い実ファイルを生成
  - 実ファイルが存在しないVFも作成可能
  - 実データファイルの生成場所を柔軟に設定可能



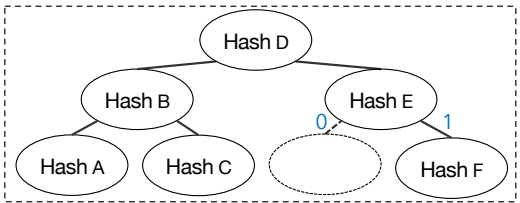
多重展開によるVFの流通促進

- VFの生成に柔軟度を与え，より広範な流通
  - VF生成を実データファイル所有者に限定しない
- 第三者がVFを発行できるメカニズム
  - 所有者および中間流通者はファイルの流通を第三者に委託可能
  - 中間流通者の情報を利用者に隠蔽
  - 所有者と中間流通者による柔軟なアクセス認可
  - 実ファイルは直接所有者から利用者へ
- 公開鍵を用いたVFのチェーン構造
  - VFの有効判定は電子署名によって実現
  - アクセス認可はチェイン形式で実ファイルまで遡及



分散ストレージ間の破損データ検出の高速化

- 分散ストレージを用いたデータ保存の課題
  - 複数の異なるストレージサービスに重複してデータを保存し破損・消失を回避
  - 従来，分散保存したデータの破損をHash値を用いて検知
  - ファイル数の増大に比例しHash値比較のコスト大
- Hash値比較演算の軽量化
  - ファイルごとのHash値をキーとした2分木を作成
  - 2分木の構造をビットマップで表現
  - ビットマップの比較によりデータ破損・消失を検出
- Hash値よりもサイズの小さいビットマップにより、狭帯域回線でも高速な破損データ検出を実現



左右の子の有無を01で表現
キーの昇順で各ノード2bitで構成

001100110100