Vol.2011-IFAT-104 No.2
2011/11/22

数値を含むテキストの類似検索が可能なフィンガープリント技術

高杰[†]　片山佳則[†]　森川郁也[†]　津田宏[†]

複数の利用者がクラウド上で情報を共有する場面が，クラウドの広がりに伴ってますます増えている．さらに，社外のクラウドに情報を預ける際に，情報の秘匿化（ハッシュ化や暗号化など)が求められている．情報を秘匿することで，情報提供先での情報漏洩やプライバシ保護などには有効である．一方で，秘匿化をすると，預けた情報を検索することは簡単にはできなくなる．これまでにある秘匿検索技術は，テキスト文書のフィンガープリントを使った秘匿類似検索や，インデックスと検索クエリを共に秘匿することで類似検索が可能な技術などである．しかし，これらの技術は，主に秘匿した文字列の類似検索ができるが，数値の特徴である範囲を指定して検索することができない．そこで，今回は数値をベースにしたフィンガープリントを提案し，そのフィンガープリントを使った数値範囲が設定できる検索システムを開発した．本システムを使うことによって，文書の原文を秘匿したまま，フィンガープリントによる数値の類似検索ができる．評価実験によると，本システムは，数値の類似検索ができない技術や，文字列ベースのフィンガープリント技術に比べ，数値の丸めがあったテストセットでの検索結果が，適合率や再現率等の指標においてより良い評価が得られた．

Fingerprint to Enable Similarity Search on Encrypted Numeric Text

Jie Gao[†]　Yoshinori Katayama[†]　Ikuya Morikawa[†]
and Hiroshi Tsuda[†]

Information sharing in cloud is increasing with widespread use of cloud. However, the data to be deposited on outside cloud should be encrypted so that confidentiality and privacy of the data is protected. Problem is that user can't search on encrypted data easily. Existing approaches including encrypting both query and retrieved data by same cryptography and key, enable search on encrypted data. These technologies also allow search on encrypted data with similar text by fuzzy matching the individual words or indices. However, users are not able to retrieve encrypted data with similar number at an assigned range. To solve the problem, we developed a new text fingerprinting algorithm which enables similarity search on encrypted numeric text. By this technology, users can retrieve encrypted data although the number in the data is rounded. In our evaluations, the proposed technology got positive search results on precision, recall and F-measure.

1. はじめに

クラウドの広がりと共に，情報をクラウドに預けてクラウド本来の特徴を生かした情報共有・活用の推進が進んでいる．

その中では，クラウドでの協業や分業における機密情報の活用が期待されている．例えば，個人が健康に関する情報をクラウドに預け，これを信頼できる公的機関などに分析・整理してもらうというような利用方法が考えられる．

こういった場面では，一部の数値情報を含む文書が共有される．例えば，医療関係では体温や血圧など患者の検査データに数値情報が不可欠である.このような情報は，共有する関係者にとっても参考になる．

一方で，セキュリティとプライバシ保護のため，これらの情報は秘匿化（ハッシュ化や暗号化など）してからクラウドに預けるのは一般的である．そうすると，セキュリティとプライバシが守られる一方で，これらの活用には制限が生じる．秘匿化情報は従来の分析アプリケーションや検索サービスでは処理できない．例えば，個人が自分の詳しい症状を公開せずに似たような人を探したい場面もある．その時に，体温や血圧等の数値上でも近い症状を持つ人のデータが参考になる可能性が高い．

我々はこのような秘匿化した情報の検索技術に着目した．そして，テキストのフィンガープリントによる類似検索技術を開発した．この技術によって，近い数値とキーワードを持つフィンガープリントを検索することができ，そのフィンガープリントから原文の持ち主を辿り着く糸口を提供できる．

フィンガープリントは，テキストの特徴を表す一種のメタ情報である．フィンガープリントから原文に戻すことができないように，中身がハッシュ関数や暗号化によって保護されている．フィンガープリントの類似検索によって，連携する各利用者が，自分の情報のセキュリティを守りつつ，情報の活用ができるようになる．

2. 従来技術と課題

これまでは，文字列の類似検索を用いたフィンガープリント技術，平文の数値の類似検索技術や暗号化した文字列の類似検索技術等の技術が開発された．それらの技術と課題について述べる．

2.1 文字列の類似検索を用いたフィンガープリント

スパム検出や情報漏洩に使われるフィンガープリント技術が開発された．これらの技術が更に 2 種類に分けられると考えられる．1 つは，ファイルのバイナリ列を分析し，ハッシュ値を取るなど，ファイルの特徴情報を抽出する技術である．特に類似するファイルを検出できる類似ハッシュがある．例えば，Kornblum[1]は，ファイルのバ

[†] 富士通研究所ソフトウェアシステム研究所
Software System Laboratories, Fujitsu Laboratories Ltd.

イナリ列を特定の式を用いて分割し，分割された各ブロックのハッシュ値を圧縮・連結したものにより Fuzzy Hash というフィンガープリントを生成する．芹田ら[2]は，Kornblum の技術を改良した類似ハッシュの生成方法を提案した．もう 1 つは，テキストの中身に基づいて，テキストの特徴的なキーワード及びそれらの関係をコーディング化したものをフィンガープリントとする．例えば，富士通のコンテンツシグネチャ技術[3]，Trend Micro 社の DataDNA[4]等があげられる．これらの技術は，テキストの原文を秘匿化したまま，原文の一部を修正（削除や移動や置換）されても類似検索することができるが，数値の類似検索はできない．

2.2 平文の数値の類似検索

吉田ら[5]は，数値を含むテキストを対象に，数値範囲指定ができるテキストの検索方式を提案した．本提案方式は，「[100..200] 円」のようなクエリに対して，100～200 円を含むテキストを見つけることができる．その他に，Google の数値範囲検索[6]や Yahoo ショッピングサイト[7]の価格範囲指定検索機能が導入されている．これらの技術は，クエリと検索対象のテキストを平文でマッチング処理を行うため，数値の比較や足し・引き算によって類似検索ができる．しかし，クエリと検索対象が秘匿されたときに，暗号化した数値の比較やソートができないため，既存の技術では，数値の範囲検索ができなくなる．

2.3 暗号化した文字列の類似検索

暗号化した情報の類似検索技術として，検索対象とクエリを同様の暗号化処理を行い，文字列間の類似度計算によって暗号化情報の類似検索方式が提案された（清水ら[8]）．Li[9]らは，インデックスとそれに指定の類似度を持つ全てのワードの集合を事前に持っておき，暗号化したクエリを集合から完全一致検索する方式を提案した．また，同研究では，ストレージと計算効率を考慮し，ワイルドカード方式による暗号化文字列のマッチングを行うことを実現した．これらの技術は，あくまで文字列の類似比較なので，数値の細かな範囲指定の類似検索は対応できない．

3. 数値ベースのフィンガープリント技術

数値の特性を利用して，秘匿化されたテキストの類似検索を行う技術は，まだ研究されていない．我々は，従来技術の課題を踏まえ数値を含むテキストの類似検索が可能なフィンガープリント技術を開発した．

そのフィンガープリントは，テキストから抽出した特徴的な数値と数値の周りのキーワードによって構成される．これらの数値とキーワードの組み合わせをハッシュ化したものをフィンガープリントとする．

フィンガープリントの生成は，大きく分けてテキスト解析等の前処理と符号化の処理になる．図 1 は，フィンガープリントの生成例を示す．それぞれの処理の詳細を以下の内容で紹介する．

3.1 正規化（前処理）

数値を含むテキストには，半角数字と全角数字，漢数字とアラビア数字，単位の違いが含まれる可能性がある．正規化処理は，これらの異なる表現を統一させることによって，検索の精度を向上することできると考えられる．例えば，全角数字「７０００」を半角数字「7000」へ，「1 万円」を「10000 円」へと変換する．本論文で取り扱うのは，全角数字から半角数字への変換のみである．漢数字とアラビア数字，そして，単位の統一は，今回は取り扱わない．

3.2 形態素解析（前処理）

テキストの特徴を表すようなキーワードや数値を抽出するため，テキストの形態素解析が必要不可欠である．形態素解析によって，テキストが分解され，品詞毎に分析することができる．我々は，オープンソースの形態素解析器と辞書を用いてテキストの解析を行った．

3.3 フィルタ（前処理）

形態素解析の結果として，テキストに含まれる名詞，動詞，形容詞等が取り出される．それらの品詞の種類の中で，テキストの特徴として利用できる品詞を選ぶための処理はフィルタである．我々は，一般名詞や固有名詞をキーワードとして，数値と合わせてマークする．

3.4 特徴素構成（前処理）

前述したように，テキストの特徴素は，数値とその周りのキーワードによって構成される．フィルタ処理によって，数値とキーワードがそれぞれ順番に並ぶようになる．特徴素は，その中の数値をまず特定し，それぞれの数値とその周りのキーワードをブロックとして構成する．N 個の数値があったときに，N 個のブロックが作られる．それらのブロックの集合が特徴素となる．

3.5 符号化

特徴素には，テキストの元情報が含まれるので，フィンガープリントとしては使えない．特徴素に含まれるすべての数値とキーワードを一般の暗号化やハッシュ化を行うことによって，元情報を隠しつつ，テキストの特徴を表すフィンガープリントを生成する．鍵を利用する場合，その鍵を提供者と検索者が共有する必要がある．

Vol.2011-IFAT-104 No.2
2011/11/22

① 原文

【患者の基本情報】
【主訴】
朝起きたら，急に体温が 38.2 度まで上昇し，発熱したような症状がみられた
【検査】
心拍数を測定した結果,85 という少し異常な値と考えられる．
【治療方針】

前処理

② 特徴素

38.2, 朝，度，体温，上昇，発熱
85, 心拍数，測定，結果，異常，値

符号化
（ハッシュ化）

③ フィンガープリント

da4e4b25e9..., a9f36f53f1..., 1cfa3d288d..., 1f2208aa5e..., 2cadce941f..., e45f66e093...
3ef815416f..., 82318fa2b7..., afc68e7e4d..., 54b7c050f2..., 623980901d..., 90013fa35b...

図 1　フィンガープリントの生成方法

数値の部分に関して，類似検索するために，元数値の有効数字や概数等の付加情報を追加することで，数値の類似検索ができるようになる．それについては，次の章で詳しく述べる．

4.　数値ベースのフィンガープリントを用いた類似検索

原文を使わずに，今回開発した数値ベースのフィンガープリントだけで類似検索する仕組みを紹介する．

4.1　概要

数値ベースのフィンガープリントを用いて，検索クエリからサーバ上のフィンガープリントを検索するシステム ANuF（Approximate Number based Fingerprint）を開発した．本システムの構成を図 2 に示す．フィンガープリントをサーバに提供するユーザとサーバ上に検索依頼するユーザをそれぞれ提供者と検索者と呼ぶ．サーバは Semi-Trusted 第三者であり，提供者と検索者の原文を知ることができないが，提供者が提供したフィンガープリント（D）と検索者が依頼したフィンガープリント（Q）のマッチングを行い，類似したフィンガープリントの提供者や提供日等の情報を検索者に返す．原文を検索者が共有したい時には，提供者の承認の上で，サーバまたは別ルートを通じて原文の共有を行う．

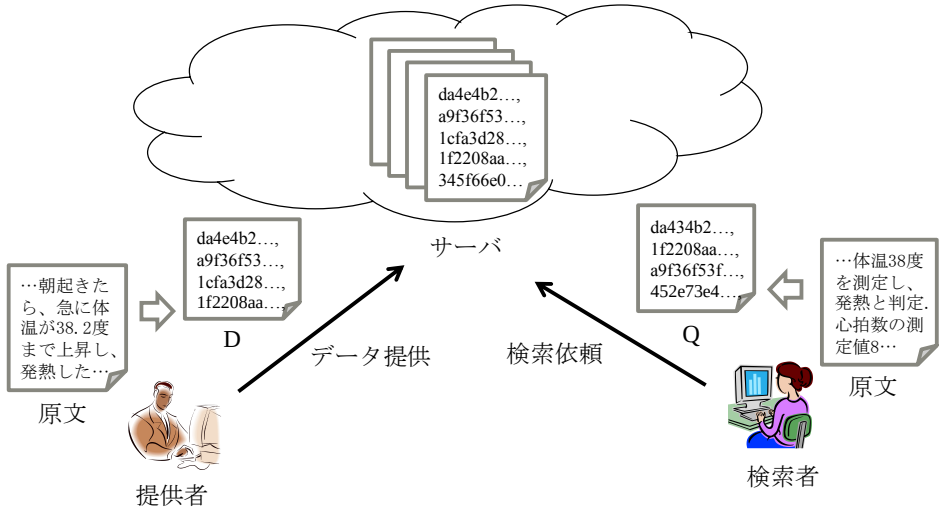

図 2　類似検索システム ANuF の構成

4.2　秘匿した数値の類似検索の原理

数値を秘匿して類似検索を行うために,フィンガープリントの数値部分を工夫する．平文でよく行うように，類似する数値は，上位の部分や概数が同じ場合が多い．例えば，38.2 と 38.0 の上位 2 桁はいずれも 38 である．我々は，数値を分離して，上位の部分を秘匿してマッチングを行い，下位の部分を無視するかまたは平文のまま足し・引き算するような方法を採用する．こういった方法は，下位の切り上げ，切り捨て，四捨五入等を含め，多く選択肢が考えられる．この論文では，下位の切り捨てと下位を平文のまま足し・引き算する 2 つの方式を紹介する．

方式１：数値を丸めて持たせる

この方式では，Q と D の数値部分に対して，数値の 1 桁，2 桁，... ，n 桁の有効数字をとった後の数値のハッシュ値を全部フィンガープリントに持たせる．例えば，図 3 の Q にある 38.2 の 1～3 桁の有効数字 3×10^1，3.8×10^1，3.82×10^1 という 3 つの数値のハッシュ値をフィンガープリントに持たせる．検索者が指定した（または，システムのディフォルトの）有効桁数で検索を行うと，その桁数での数値のハッシュ値の同じものが類似と判定される．例えば，指定された有効桁数が 2 桁のときに，図 3 の D1 と D2 は Q の数値に類似するが，D3 は類似しない．指定された有効桁数が 3 の場合は，D1 の数値しか類似しない．

Vol.2011-IFAT-104 No.2
2011/11/22

Q	D1			D2			D3		
38.2	**38.2**	**有効桁数**	**結果**	**38**	**有効桁数**	**結果**	**39**	**有効桁数**	**結果**
3×10^1	3×10^1	1	≒	3×10^1	1	≒	3×10^1	1	≒
3.8×10^1	3.8×10^1	2	≒	3.8×10^1	2	≒	3.9×10^1	2	≠
3.82×10^1	3.82×10^1	3	≒	3.80×10^1	3	≠	3.90×10^1	3	≠

図 3　類似検索の方式 1

方式２　数値の概数とそれとの差分を持たせる

この方式では，D の数値の前後一定範囲内の概数及び概数と元数値の差分を求め，概数のハッシュ値と差分をフィンガープリントに持たせる．例えば，図 4 に書いた例のように，D の数値 38.2 が前後の概数を取るときに，30 と 40 が考えられる．概数のハッシュ値（H（30），H（40））と概数と元数値の差分（8.2 と-1.8）を平文で提供することで，元数値の上位を隠しながら，下位の数値による比較や足し・引き算が可能となる．一方で，Q の数値の概数と元数値と概数の差分を求め，概数のハッシュ値と差分及び数値範囲の情報が含まれる．図 4 で示したように，39.1 を 30 と 9.1 に分け，30 のハッシュ値，9.1 及び数値範囲 1 をフィンガープリントに入れる．

検索を行うときに，Q と D の数値部分のハッシュ値（H（30）と H（30），H（40））を比較し，一致したもの（H（30））に関しては更に差分部分の差(9.1－8.2)を計算する．その結果 0.9 が検索者が指定した範囲 1 より小さければ，数値の部分が近似と判断する．

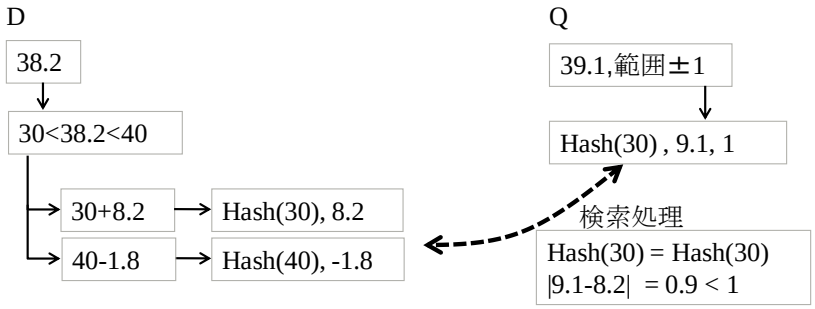

図 4　類似検索の方式 2

4.3　フィンガープリントの構成

前節の類似検索方式に従って，フィンガープリントの数値部分を構成する．図 5 はフィンガープリントの構成を示す．1 つの数値（Num）とその周りのキーワード（KW）を 1 つのブロック（Block）と呼ぶ．数値には，更に類似検索のための付加情報（Aux）を含む．数値とキーワードの全部と付加情報の一部または全部がハッシュ化されたものである．

$Block_1$: $Num_1(Aux_{1.1}, Aux_{1.2}, ..., Aux_{1.X})$, $KW_{1.1}, KW_{1.2}, ..., KW_{1.Y}$
$Block_2$: Num_2 $(Aux_{2.1}, Aux_{2.2}, ..., Aux_{2.X})$, $KW_{2.1}, KW_{2.2}, ..., KW_{2.Y}$
...
$Block_N$: Num_N $(Aux_{N.1}, Aux_{N.2}, ..., Aux_{N.X})$, $KW_{N.1}, KW_{N.2}, ..., KW_{N.Y}$

図 5　フィンガープリントの構成

4.4　類似度計算

Q と D の類似度は, Q と D の共通の類似数値が属するブロック間の共通キーワードの数と Q のすべてのキーワードの数の比率である. Q と D に含まれる同じ数値のブロックが統合され，ブロック内の数値とそれぞれのキーワードのペアは唯一である．数式 1 は，Q と D の類似度の計算式を示す．

この式の特徴は，

・ 類似度の範囲は 0（低い）～1（高い）である．
・ Q と D に共通する数値がなければ，類似度が 0 となる
・ Q と D に共通する数値があっても，その数値の周りに共通のキーワードがなければ，類似度が 0 となる ．
・Q と D が同じフィンガープリントの場合，類似度が 1 となる．

数式 1　$$Similarity(Q,D) = \sum_{i,j=0}^{N_Q,N_D} \frac{Sim(Num_{Q.i}, Num_{D.j}) \times (B_{Q.i} \cap B_{D.j})}{Win\_Size \times N_Q}$$

ただし，
$\mathbf{N_Q}$，$\mathbf{N_D}$ は Q と D のブロックの数，
Win_Size は 1 つのブロックに含まれるキーワードの数，
$\mathbf{B_{Q.i}}$ と $\mathbf{B_{D.j}}$ は Q と D のブロック，
$\mathbf{Num_{Q.i}}$ と $\mathbf{Num_{D.j}}$ は Q と D のブロックに含まれる数値，
$\mathbf{B_{Q.i}} \cap \mathbf{B_{D.j}}$ はブロック $B_{Q.i}$ と $B_{D.j}$ の共通のキーワード数であり, Sim()は 数値の類似度判定式で，次の数式 2 で定義される．数値が類似するかしないかは，採用された類似検索の方式によって決められる．

数式 2　$Sim(Num_{Q.i}, Num_{D.j}) = \begin{cases} 1 & Num_{Q.i} \text{ と } Num_{D.j} \text{ が類似する} \\ 0 & Num_{Q.i} \text{ と } Num_{D.j} \text{ が類似しない} \end{cases}$

類似度計算の例として，4.2 の類似検索方式 1 を使って図 1 の②の特徴素に付加情報を追加して生成したフィンガープリントを Q として，図 6 の特徴素から生成したフィンガープリントを D として，Q と D の類似度を計算する（方式 1 における有効桁数は 1 と指定する）．Q と D の数値部分の 1 桁に着目したときに，「3×10^1」と「8×10^1」の 2 つの類似数値を特定できる．「3×10^1」が対応する Q と D のブロックの共通キーワードは，「度」，「体温」，「発熱」の 3 つである．それを「8×10^1」の対応ブロックの共通キーワード数 2 を合わせると，Q と D の共通キーワード数は，5 となる．更に，ブロックサイズ 5 と Q のブロック数 2 を数式 1 に代入すると，類似度は 5/(2×5) = 0.5 となる．（実際の共通の数値とキーワードの特定は，全部ハッシュ値のまま行われる．）

38.2 (3×10^1, 3.8×10^1, 3.82×10^1)，体温，度，発熱，測定，判定
83 (8×10^1, 8.3×10^1, 8.30×10^1)，　心拍数，測定，正常，症状，高
3(3.0×10^0, 3.0×10^0, 3.00×10^0)，　軽度，発熱，薬，経過，再診

図 6　特徴素

4.5 類似検索の手順

提供者，検索者とサーバが協力し，4.2 で述べたいずれの方式を共通方式として決め，それに沿ってフィンガープリントを提供，検索する．本節は，提供者，検索者とサーバのそれぞれの処理について紹介する．

S1：フィンガープリントの提供

提供者は自分が抱えている文書のフィンガープリントをクラウドに預ける．フィンガープリントの生成にあたって，共通方式に従って付加情報をフィンガープリントに持たせる．サーバは預かったフィンガープリント，及び提供者や提供日の情報をフィンガープリント DB で管理する．

S2：フィンガープリントの検索

検索側は検索したい文書のフィンガープリントをサーバに送って検索依頼する．提供側と同じように，フィンガープリントは共通方式に沿って生成したものである．

S3：サーバでの検索処理

サーバは，フィンガープリント DB のすべてフィンガープリントを検索依頼されたフィンガープリントの類似度を計算する．類似したものを類似度順にソートし，類似度の高いフィンガープリントの提供者情報などを検索結果として検索者に返す．検索結果を絞るためには，ある指定の類似度以上のものや上位 N 個のものなどの方法が考えられる．図 7 は，検索者に返す検索結果の例を示す．

検索依頼されたフィンガープリントに類似するものは以下の 3 件あった．

フィンガープリント No.	提供者	提供日	類似度
F001	Alice	2011/1/1	0.8
F101	Bob	2011/8/2	0.6
F051	Charlie	2011/3/3	0.5

図 7　検索結果の例

5. 評価実験

フィンガープリントを用いた類似検索システム ANuF の性能を検証するための評価実験について述べる．

5.1 概要

今回の評価実験では，2 つの比較対象システムと ANuF を用いて，ウェブ記事を元に作成したテストセットから類似するフィンガープリントを検索させた．評価指標としては，情報検索でよく使われる適合率，再現率と F 値を用いた．ANuF の数値の類似判定は，方式 1 を採用した．そして，数値の有効桁数を 1 桁とした．

5.2 比較対象

ANuF の比較対象として，次のようなシステムを開発した．

比較対象 A： 数値ベースで，数値の近似検索ができないシステム

比較対象 A は，フィンガープリントの生成方法と類似度の計算式に関して，ANuF と同じであるが，数値に関する付加情報を持たない．そのため，数値が完全一致ではないと，マッチングさせない．

比較対象 B： キーワードベースで数値の近似検索ができないシステム

数値とキーワードを区別せずに，テキストに含まれるすべての数値とキーワードを抽出する．そして，それらの数値とキーワードの中で，共通の数値とキーワードが占める比率を類似度とする．

5.3 テストセットの作成

今回の評価実験のために，複数のサイトからおよそ 4600 件の記事を収集した．その中で，ある出来事（テーマ）について，異なるサイトで掲載した記事を類似文書（検索時の正解文書）とした．今回の実験では，数値を扱う 3 つのテーマについての記事を採用した．さらに，それぞれのテーマに関する類似文書を元に，数字の部分の編集

によって，100 個のテストセットを作成した（図 8 参照）．今回は，主に有効数字 3 桁以上の数値の最下位の 2 桁に対して，切り上げ，切り捨て，四捨五入とランダム化の丸め処理を行った．作成されたテストセットは提供者の原文として，それぞれの検索システムでフィンガープリントが抽出され，サーバに提供された．

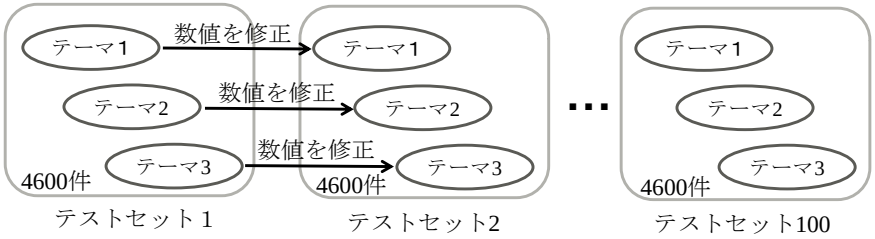

図 8　テストセットの作成

5.4　評価方法

それぞれのテーマについて，関連記事を元にテキストを作成した．それらのテキストのフィンガープリントを検索クエリ（Q）として，各検索システムを使ってサーバに提供されたフィンガープリント（D）を検索する．検索結果を，適合率，再現率，F 値の 3 つの指標で評価する．

適合率　検索結果の中で正解文書が占める比率
再現率　検索結果に含まれる正解文書とテストセットにある正解文書の全体の比率
F 値　適合率と再現率の調和平均値

検索結果を絞るために，類似度でソートした類似フィンガープリントの上位 10 個を検索結果とする．それぞれのテーマの正解文書数は，いずれも 10 より小さい．

5.5　実験結果

表 1，　2 ，3 に，テーマ別の各システムで検索した結果の適合率，再現率，F 値の平均値をまとめた．（各数値は小数点第 2 位未満を四捨五入した．）

表 1　適合率の平均値

	ANuF	比較対象 A	比較対象 B
テーマ 1	0.7	0.7	0.6
テーマ 2	0.5	0.41	0.21
テーマ 3	0.6	0.42	0.4
平均	**0.6**	0.51	0.4

表 2　再現率の平均値

	ANuF	比較対象 A	比較対象 B
テーマ 1	0.88	0.88	0.75
テーマ 2	0.63	0.52	0.26
テーマ 3	0.86	0.59	0.57
平均	**0.79**	0.66	0.53

表 3　F 値の平均値

	ANuF	比較対象 A	比較対象 B
テーマ 1	0.78	0.78	0.67
テーマ 2	0.56	0.46	0.23
テーマ 3	0.71	0.49	0.48
平均	**0.68**	0.58	0.46

この 3 つの表から，ANuF が他の 2 つのツールよりも，適合率，再現率と F が高いことが分かる．それは，テキストの数値が一部丸められていても，ANuF はフィンガープリントに含まれる数値の付加情報によって，数値の類似検索ができるからである．

テーマによっては，各指標の変動があると見られる．それは，Q と D に含まれる数値の数や，丸められた数値の数による違いと考えられる．

表 4 は修正前の D（3 テーマ×1 テストセット）と修正後の D（3 テーマ×99 テストセット）の各指標の平均値を示す．（各数値は小数点第 2 位未満を四捨五入した．）

表 4　数値の修正前後の各指標の比較

	適合率		再現率		F 値	
	修正前	修正後	修正前	修正後	修正前	修正後
ANuF	0.6	0.6	0.79	0.79	0.68	0.68
比較対象 A	0.63	0.51	0.83	0.66	0.72	0.58
比較対象 B	0.43	0.4	0.57	0.53	0.49	0.46

表 4 からは，ANuF は数値の修正前後，各指標の値は変わっていないことが分かる．それは，数値が修正されても，最上位 1 桁の値が変わっていないので，ANuF の検索結果に影響が殆どない．比較対象 A と B は，修正前後の各指標の値が下がったが，特

に比較対象 B の下がり幅が大きい. その原因は検索対象 A が扱うフィンガープリントは数値ベースなので，数値が修正によって完全一致しなくなる数値が多く出ていたので，検索結果が多く変わった．それに対して，検索対象 B の場合は，数値ベースではないので，一部の数値が変わっても，大多数を占めるキーワードが変わっていないので，検索結果に変化を殆どもたらしていないと考えられる．

6. 応用

個人が健康に関する情報をクラウドに預け，これを信頼できる公的機関などに分析・整理してもらうというような利用方法が考えられる．この場合，それぞれの提供者は情報を秘匿化しているため，他の利用者はどのような情報を提供しているかは分からない．しかし，健康情報や公的機関の分析結果を参照したい個人は自分に似た症状を処置している他の提供者とは情報交換したい場面もある．この際に，ここで提案する活用方法を用いることで，クラウド上では，互いの提供内容が漏れることなく，クラウドで管理されている秘匿化情報を用いて，関連する提供者同士が安全に情報交換可能になる．

また，教育機関などで，レポートに盗用・盗作が含まれるかどうかをチェックするような機能が必要とされている．本技術は，文書の類似度チェック以外に，数値の微妙な修正によって盗用・盗作をごまかそうとすることも検知できる．

その他に，添付送信したファイルは，受信者がどのように操作されたかを追跡するトレーサビリティ技術[10]において，ファイルが修正されたり別のファイルとマージされたりしたときに，更新後のファイルと元ファイルのフィンガープリントの類似度計算によって，操作イベントを検知することができる．また，修正後のファイルを外部へ誤送信しようとするときに，元ファイルとの類似度が誤送信をブロックする根拠にもなる．

7. まとめ

数値を含むテキストの類似検索が可能なフィンガープリント技術及びそれを用いたテキストの類似検索システム ANuF を開発した．本技術を使うことによって，原文を使わずに，フィンガープリントだけで類似検索が可能となる．評価実験により，ANuF が，数値の改変に耐性を持つことが検証された．本技術は，クラウドでの機密情報活用に役立つと期待される．今後は，フィンガープリントの圧縮等に取り組む予定である．

参考文献

1) Jesse D. Kornblum: Identifying almost identical files using context triggered piecewise hashing, Proceedings of the Digital Forensic Workshop, pp. 91-97 (2006)
2) 芹田進，藤井康広，甲斐賢，村上隆夫，本多義則: ファイル伸縮に耐性のある類似ハッシュ算出方式の考察，電子情報通信学会技術研究報告, Vol. 110, No. 281, pp. 31-36 (2010)
3) 竹林知善，津田宏，長谷部高行，益岡竜介: 情報漏えい防止セキュリティ技術開発への取組み, Fujitsu, No. 60, pp. 444-450 (2009)
4) http://jp.trendmicro.com/jp/products/enterprise/tmdlp/technology/
5) 吉田稔，佐藤一誠，中川裕志，寺田昭: 接尾辞配列とディリクレ過程混合モデルを用いたテキスト中の数値表現マイニング，研究報告情報学基礎（FI）, Vol. 2009-FI-96, No. 4, pp. 1-8 (2009)
6) http://www.google.com
7) http://shopping.yahoo.co.jp/
8) 清水將吾，權娟大: DAS モデルにおける安全な類似文字列検索方式の提案，日本データベース学会論文誌, Vol. 7, No. 1, pp. 13-18 (2008)
9) Jin Li, Qian Wang, Cong Wang, Ning Cao, Kui Ren, and Wenjing Lou:. Fuzzy keyword search over encrypted data in cloud computing, Proceedings of the 29th conference on Information communications (INFOCOM'10), pp. 441-445 (2010)
10) 高杰，園田俊浩，片山佳則，津田宏: メール添付ファイルのトレースシステムの試作，第 72 回情報処理学会全国大会講演論文集, No. 3, pp. 585-586 (2010)