D-041

サービス利用状況に着目した情報格差観測データ流通基盤の設計

Design of distribution foundation for digital devide observation data focusing on states of service utilization

岩田 翔汰 † 中平 勝子 † 北島 宗雄 †
Shota Iwata Katsuko T. Nakahira Muneo Kitajima

1 はじめに

現代のインターネットにおける情報格差（以降，情報格差）は，基盤，利用，効用それぞれの行為に対して発生しうる．本稿では，利用情報格差に着目し，その動向調査を継続的に行うために必要な情報格差観測データの分散管理・提供手法についての検討を行う．利用における情報格差とは，情報通信技術の利用の度合いから生じる情報格差のことである．今日，非常に速いスピードで様々なサービスが生み出され，普及・衰退していくが，それらのサービスの利用における情報格差をリアルタイムに把握することは困難である．利用情報格差は，既存サービスのみならず新規に生じたサービスへのアクセス困難度も問題となるため，常に多地点での継続的観測および情報共有が必要となる．そのための適切な観測データ共有のスキームについて言及する．

安全・安心かつユビキタスなネットワークの実現のためには，すべてのカントリー・コード・トップ・レベル・ドメイン（ccTLD）管理者が一定の水準を満たすカントリードメイン管理（カントリードメイン・ガバナンス，以下 CDG）を行う必要があり，CDG の実態を 3 つの基準（アクセス，言語多様性，安心と信頼）から多面的に評価する指標として CDG Index が開発されている [1]．また，情報の流通を可能にしている ICT ネットワークを e ネットワークと呼び，情報コンテンツがどのようにして観測される姿で存在し得るかという視点で複雑な社会現象であるネットワーク生態を解明するという研究が存在する [2]．

伊藤らは，情報格差分析に利用するクロールデータや統計情報などを 1 次処理したデータを研究者間で流通させるために必要なデータ形式 Information Trade Handling Format（以下，ITHF）および，情報格差を分析するシステムを提案している [3]．本稿では，伊藤らの開発したデータ形式による情報格差観測データの保管・提供方法の設計を行う．

2 情報格差観測の流れ

大規模なデータは，それ自体の流通によって，その分野や関連分野が発展したり，規模が拡大するという特徴がある．流通したデータは研究者ごとに取得され，分析が行われる．ビッグデータ分析には「データ収集」「情報抽出とクリーニング」「データ統合，データ集約，データ表現」「モデル化と分析」「解釈」「不均質性」の 6 つの特徴がある [4]．情報格差観測データについても，これらの特徴にしたがって分析を行う．

表 1 に CDG 評価指標体系を示す．これらの指標は，基本的に，オープンな情報源から，継続的に，客観性をもって導出することのできる指標群である．LLPP や RPLL などの指標については Web ページのクローリングによって得られるデータや公的機関が発行している統計情報によって計算できる．CDG 評価指標体系の各指標を e ネットワークの枠組みの中でとらえることによって，情報格差の解明につながるのではないかと考えられる． 表 1 を踏まえた情報格差観測データの収集・分析手順について述べる．情報格差観測データは，インターネット上における種々のデータ，例えば，URL のリンク関係や Web ページで使用される言葉やマルチメディアファイル，拡張すれば，Web のみならずインターネット上における全てのパケットに関するデータなど，多くの情報を扱う．この意味において情報格差観測は一種のビッグデータと捉えることができ，その流通には先に述べた 6 つの特徴を捉えながら格差観測や観測データの流通を行う必要がある．その枠組みを図 1 に示す．

Web ページのクローリングを行うことで，URL やリンク情報，コンテンツ等の様々な情報が得られる（データ収集）．情報格差の観測に直接使用しないデータを取り除く一次処理（クリーニング）を施し，また，地理情報の取得や言語解析を行うことで，URL データ，LINK データ，サーバ位置情報，言語解析データといった情報格差データが得られる（情報抽出）．また，公的機関が発行する統計データも情報格差観測に用いる必要があるため，他のデータと合わせて（データ統合，データ集約），分析に適した形に成型する（データ表現，不均質性の解決）．これらのデータを総合的に分析し，CGI Index のような情報格差に関する指標を計算することで（分析，モデル化），情報格差の解明（解釈）につながるのではないかと考えられる．

これらの特徴のうち，データのクリーニングや情報抽出，分析については言語天文台プロジェクト [5] の中で，データ統合・集約は ITHF によってそれぞれ行われており，データのモデル化および解釈は e ネットワーク枠組みの中で行うことが可能である．これら一連の流れは，1 機関のみで行われる情報格差観測でも対応可能である．しかし，データの不均質性については，1 機関のみの格差観測では部分観測しかできないこともあり得る．特にクローリングによるデータの収集は，インターネット全体からみると一部にすぎず，1 機関のクロールだけでインターネット全体を網羅するのは不可能である．情報格差の観測を実情に近づけるためにはデータ量の増大が要求される．そこで，クロールデータを収集している研究者同士がデータを

† 長岡技術科学大学

表 1　CDG 評価指標体系 (Country Domain Governance Index)

	Category	Indicator	Description
アクセス	Accessible And Affordable	①RPDM	ドメインの相対価格（月間所得との比率） ドメイン管理者が公表する価格と公的機関が発行する所得に関するデータから計算
		②RPDG	ドメインの相対価格（世界平均価格との比率） ドメイン管理者が公表する価格から計算
		③NDPP	人口当たりのドメイン発行数 ドメイン管理者が公表するドメイン数と公的機関が発行する人口データから計算
	Openness of Network	④RSLO	サーバのドメイン外設置比率 URL/LINK データの HDU の位置情報から計算
		⑤NOLM	国際ニュースメディアへのリンク数 LINK データの内国際ニュースメディアに向けて張られたリンクの数
言語多様性	IDN service	⑥NIDN	国際ドメイン名（IDN）発行数 URL データの内国際化ドメインになっているものの数
	Local Language Use	⑦LLPP	人口当たり現地言語ページ数 LANG データと公的機関が発行する人口データから計算
		⑧RPLL	現地言語ページ比率 LANG データの公用語と現地言語の比率
		⑨LDLI	リーバーソン指標で測った言語多様性 公的機関が発行する言語別話者数のデータから計算
安全と信頼	SSR	⑩SSMO	スパムメール発信源比率 スパムメールフィルタによってはじかれたメールの IP アドレスの位置情報から求める
		⑪RDRA	ドメインの匿名登録者比率 公的機関が発行する統計データから計算
	Trusted contents	⑫ADRP	紛争解決手続きの利用可能性 ドメインごとに記載されている紛争解決手続きを確認することで求める

共有し，流通させることが有効な解決策となる．本稿ではこのデータ流通の方法について検討を行う．データ流通方法について述べる前に，ITHF の概要と構造，ITHF の生成・分析を行うシステムについて説明する．

クロールデータや統計データに一次処理を行い，複数のデータを共通して扱えるようにしたフォーマットが ITHF である．ITHF はヘッダとデータを併せ持つブロック Header Data Unit（以下，HDU）を複数保持するデータ格納構造をしている．ITHF には情報格差観測データの種類の数だけ HDU が格納される．ITHF の設計時には，表 1 に従い，必要とされる情報格差データは，

- STAT データ：クロールを実施した国に関する統計データ
- URL データ：クロールによって得られた Web ページの URL
- LINK データ：URL に含まれるハイパーリンク
- LANG データ：Web ページで使用されている言語情報
- LOC データ：URL のサーバが設置されている所在地情報

の 5 つが想定されている．これらのいずれか，または全てが HDU として ITHF に格納される．情報格差に関する新たな指標が提案された場合，その指標に関する分析に必要となるデータの HDU を定義し，HDU を生成し，分析ツールにモジュールを追加することで，新たな指標に関しても分析が行える．

3　ITHF の管理・流通

図 2 は，ITHF の管理・提供方法を模式的にあらわしたものである．この内容について，以下で説明する．

ITHF の管理：ITHF そのものはファイルサイズが大きいため，1 箇所に集約しようとすると，大規模なストレージと高速なネットワーク回線が要求される．そこで，ITHF を集約するのではなく，ITHF へのリンクを含む ITHF の概要を示したファイル（以下，ITHF 概要ファイル）を集約して管理する分散管理手法を考える．

伊藤らが開発した ITHF の分析システムには，統計データやクローリング結果から ITHF を生成する機能が備わっている．このシステムを，ITHF の生成と同時に ITHF 概要ファイルの生成が行われるように変更する．

この時，同時に生成する ITHF 概要ファイルには以下の項目を記す．

- COUNTRY：ITHF がどの国を対象とした情報格差観測データなのかを示す．
- HEADSIZE，DATASIZE：ITHF のヘッダー部分とデータ部分のサイズを示す．これらの値の合計から，ITHF 全体のファイルサイズを求める．
- DATE：HDU を作成した日付
- TIMESYS：HDU 作成地点のタイムゾーンと協定世界時との差
- AUTHOR：HDU を作成・更新した人の名前
- ORG：HDU 作成者が所属する組織名
- DATEORIGINAL：統計データが発行された日付やクロールの実施日
- TIMESYSORIGINAL：統計データの発行地点やクロール

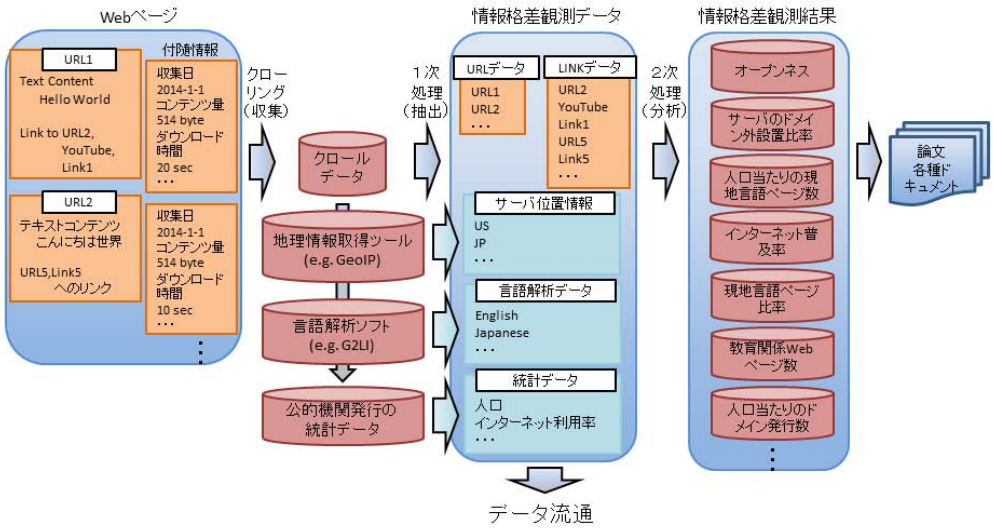

図 1　「利用の格差」観測手順

表 2　ITHF 概要ファイルのデータベースのイメージ

ID	URL	COUNTRY (ccLTD)	SIZE (Byte)	HDU No	EXTNAME	DATE-TIMESYS	AUTHOR	DATE-TIMESYS ORIGINAL	AUTHOR ORIGINAL
001	http://abc.jp	jp	1234567890	HDU0	STAT	2015-08-01 00:00:00 +0000	Iwata	2015-07-01 12:00:00 +0900	UN
				HDU1	URL	2015-08-01 01:00:00 +0900	Iwata	2015-07-15 12:00:00 +0900	Nakahira
				HDU2	LINK	2015-08-01 02:00:00 +0900	Iwata		Nakahira
002	http://aaa.kr	kr	2345678901	HDU0	URL	2015-09-01 13:00:00 +0900	Iwata	2015-08-15 11:00:00 +0000	Iwata
				HDU1	LINK	2015-09-01 14:00:00 +0900	Iwata		Iwata

の実施地点のタイムゾーンと協定世界時との差

- AUTHORORIGINAL：統計データを公開した人の名前やクロールを行った人の名前
- ORGORIGINAL：統計データの発行組織やクロールの実施組織
- EXTNAME：それぞれのユニットにはどのようなデータが記録されているのか（例，EXTNAME が LANG なら Web ページの言語情報）

ITHF を提供するユーザは，これらの項目を記した ITHF 概要ファイルをサーバにアップロードする．この時に，対応する ITHF の URI を指定する．アップロードを受け付けるサーバは，指定された URI に存在する ITHF のファイルサイズと，ITHF 概要ファイルに書かれている HEADSIZE・DATASIZE の合計値を比較し，ITHF と ITHF 概要ファイルが正しく対応していることを確認し，アップロードを受け付ける．ITHF 概要ファイルのアップロードが行われた後，ITHF 概要ファイルの内容が自動的にデータベース化される．表 2 は ITHF 概要ファイルデータベースのイメージである．データベース化と同時に ITHF 一覧データの生成も行う．ITHF 一覧データは独自に拡張した XML ファイルとし，ITHF 概要データと同じ内容が記述されている．新しい ITHF 概要データがアップロードされた場合に，ITHF 一覧データに次々と内容を追記していく．

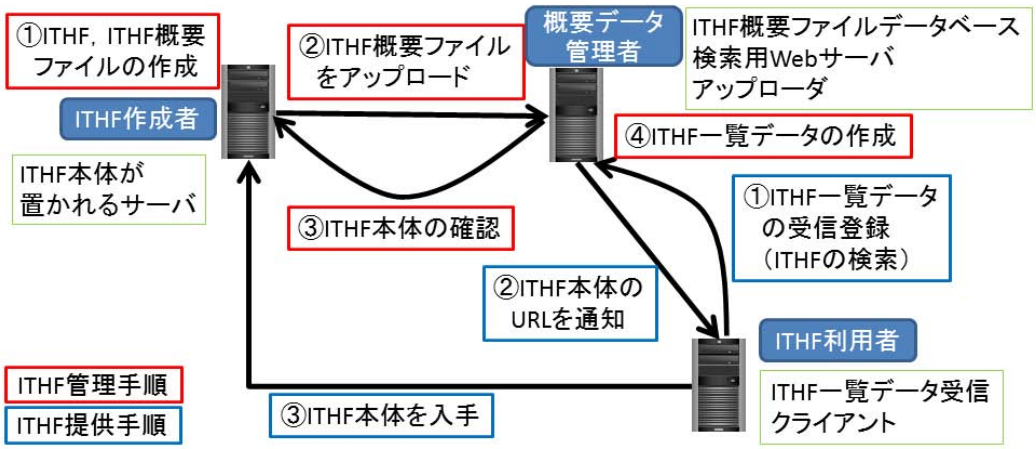

図 2　ITHF の管理・提供方法のイメージ

ITHF の提供：ITHF の提供方法として 2 つの方法が考えられる．ITHF 概要ファイルのデータベースを検索するインタフェースを提供し，データベースに記録された ITHF 概要ファイルの URI から ITHF へアクセスする方法，もう 1 つは，RSS や Atom のように，ITHF 概要ファイルデータベースに新しい概要ファイルがアップデートされると，ITHF 一覧データが更新され，リンクをたどって ITHF にアクセスする方法である．

ITHF を利用したいユーザは，ITHF 概要ファイルを管理するサーバで検索し，ITHF にアクセスを行う．もしくは，ITHF 一覧データを RSS リーダーのようなソフトウェア（や Web ブラウザ）に登録し，ユーザに届いた更新通知の URI から ITHF にアクセスする．RSS や Atom と同様に，ITHF 一覧データには個人認証の仕組みがないため，ITHF 一覧データの URL を暗号化して利用登録を申し込んだ者に提供し，IP アドレスのログを取るというような運用も考えることができる．ただし，ITHF 一覧データの URI が外部に公開されてしまうと誰でもアクセスが可能になるため，アクセスに制限をかける場合には，ITHF 一覧データ自体の暗号化を検討する必要がある．

4　技術要素

前章で説明したデータ管理・提供スキームを実装するために必要な技術要素について具体的に述べる．

ITHF を作成し所有する研究者等は，リーダー（もしくは Web ブラウザ）が通常の方法でアクセスできるプロトコル（HTTP など）で ITHF をアクセス可能な状態に保つことが求められる．ITHF 概要ファイルがアップロードされるサーバやデータベースが保存されるサーバはアクセス回数が多くなり，とても重要な役割を果たすため，アクセス速度が高速でかつ安全に動作する場所に安定的に動作するように設置されるべきである．また，ハードウェアの多重化も行う必要がある．アップロードを受け付けるサーバは Perl や PHP などで書かれた汎用の CGI ベースのアップローダを利用することで実装可能である．ITHF 概要ファイルのデーターベースは PostgreSQL などのフリーソフトを利用することで実装可能であるが，外部から不正にアクセスされないように正しいアクセス権限の設定が必要となる．ITHF 一覧データについては，Atom や RSS の形式を独自に拡張するため，専用のリーダーを作成しなければならなくなるということも考慮する必要がある．

5　まとめと今後の課題

本稿では分散管理・提供を通して観測データの流通を円滑に行うための仕組みを設計した．問題点として，Atom や RSS の形式をベースに情報格差一覧データを配布するようにしたため，利用者の認証ができない．ITHF データの利用の仕方が問題となった場合に利用者の追跡が出来ないため，データの暗号化や認証といった内容についてさらに理解を深めたい．

参考文献

[1] 三上喜貴：「カントリードメインの脆弱性監視と対策」研究開発実施終了報告書（科学技術振興機構・社会技術研究開発事業研究開発プログラム「ユビキタス社会のガバナンス」/研究代表者：三上喜貴）(2011)

[2] 中平勝子，北島宗雄：人の営みとして形成される e ネットワークのダイナミクスを解明するための枠組み，ARG WI2，No.1，pp.47-48，2012

[3] 伊藤公：情報格差観測データ流通のためのファイルシステムの開発・評価，長岡技術科学大学修士論文，2015 年 1 月．

[4] H.V. Jagadlish et al.: Big data and its technical challenges, Communications of the ACM Volume 57 Issue 7, July 2014

[5] 三上喜貴，中平勝子，児玉茂昭：言語天文台からみた世界の情報格差，慶應義塾大学出版会，2014