5ZD-06

# 大容量データ流通のためのファイルシステムの開発・評価

伊藤 公　　中平 勝子　　三上 喜貴

長岡技術科学大学

## 1. はじめに

本研究では，クロールによって収集した Web ページデータ（クロールデータ）を研究者間での研究促進を図るために設計した Information Trade Handle Format(ITHF)[1]を用いて，インターネット上のデジタル・デバイドを分析するシステムの開発および検証を行う．

近年，情報通信技術やインターネットサービス（例えば，ソーシャルネットワークサービス(SNS)）の発展と普及に伴って日々大量の Web ページが生成され，その Web ページを他ユーザと共有するソーシャルブックマークの増加により大量の Web リンクが継続的に生成されている[2]．これらの Web ページをクロールにより収集し言語天文台[3]における Web ページ上のリンク解析や言語解析を行うに当たり，以前と比べて処理すべきデータが増加している．収集すべき Web ページが増加する一方，1 機関のクロールデータの収集規模には多くても数 TB という限界がある．それに対し，インターネット上を流通するコンテンツ量（テキスト，画像，動画・音声ファイル等）は 2009 年時点で 7PB(JP ドメイン上)であり[4]，現在ではより大量のコンテンツが流通していると予想される．

インターネット上のデジタル・デバイドの実態把握には，多くのクロールデータを使い分析をすることが求められる．そのため，分析対象となる源データを如何に増やすか模索する必要が生まれてきた．

そこで我々は，異なるフォーマットで収集されるクロールデータを可能な限りそのままに近い形で流通させ，研究者間で共有するための汎用性の高いファイルフォーマット(ITHF)を設計し，利用することを提案した．ITHF では，クロールデータから抽出した Web ページに関わる一次処理済みデータ，統計情報等を一つのファイルに統合する．これにより，他機関が収集したクロールデータを活用でき分析精度の向上や研究コストの削減等が期待される．この環境整備には，ITHF の検証[5]や ITHF を使った分析システムを構築が必要となる．

本稿では，ITHF を使った分析システムの開発を行い，デジタル・デバイドを従来手法で分析した結果と開発したシステムで分析した結果の差異，処理時間の変化について検証を行う．

## 2.開発したシステムの概要

本研究で開発するシステムには，次の機能が求められる(図 1)．

(1)の機能は，ITHF を流通させるため，ITHF を作成するために必要となる．

(2)の機能には，ITHF を設計する際に参考にした天文分野で利用される FITS[6]の分析に使用するソフトウェア IRAF[7]の構想を取り入れる．IRAF では，分析を行う際，分析目的に合わせて必要な”パッケージ”（分析目的の処理タスクをまとめたかたまり）を取り入れる事により，システムに依存せず多様な分析を行うことが可能である．

この”パッケージ”をデジタル・デバイドの分析指標と対応させ，”パッケージ”を取り入れる事により多種多様なデジタル・デバイドの分析が出来るよう開発を行った．システムでは一つの分析指標を算出する関数を一つの”パッケージ”として，この関数をまとめた実行ファイルを用意した．今後新しい分析指標が開発された場合には，その分析処理タスクを関数にして実行ファイルをシステムに加えることで，システムに変更を加える事なく分析が可能となる．

## 3. 開発したシステムの検証

検証で行うデジタル・デバイドの分析指標には，以下の指標を利用する．

・クロールデータから抽出した URL に含まれる LINK を使った分析指標(OPENNESS[2])

・Web ページのサーバ設置情報に基づく分析指標(サーバのドメイン外設置比率，RBRLL[8])

これらの分析処理タスクを”パッケージ”としてまとめ，システムに取り組み分析を行う．このシステムを利用した結果，このシステムを利用する以前から分析に利用していた Perl を使ったプログラムで算出した結果に差異がないか確認を行う．また，その際の処理時間の変化についても確認を行う．それぞれの分析方法と各分析指標での処理時間の結果を表 1 に示す．$T_o$は

File System for Circulating Massive Data: Development and Evaluation
Akira Ito, Katsuko T. Nakahira, Yoshiki Mikami
Nagaoka University of Technology

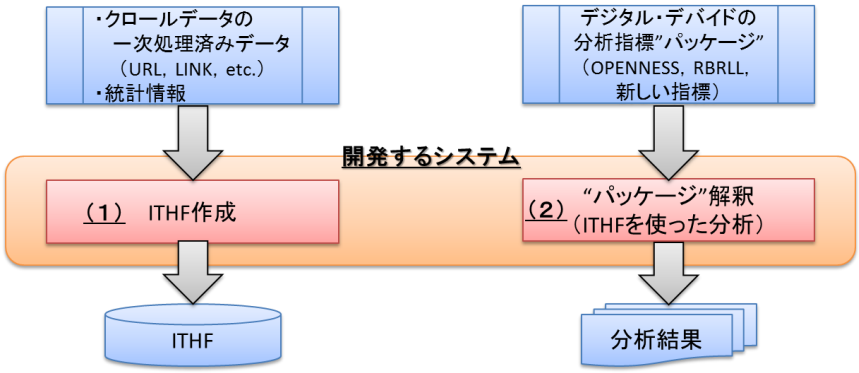


図 1. 開発するシステムの概要図

表 1. デジタル・デバイド分析に要した処理時間

| 処理体系 | データ 1 | | データ 2 | |
|---|---|---|---|---|
| | $T_{o1}$[s] | $T_{r1}$[s] | $T_{o2}$ [s] | $T_{r2}$ [s] |
| Perl | 0.015198 | 0.004167 | 477.302822 | 14.376536 |
| システム | 0.000530 | 0.003608 | 3039.777165 | 7.498770 |

OPENNESS, $T_r$ は RBRLL の処理時間を示す. ITHF を作成するのに用いたクロールデータは, 1MB と 22GB のデータ（データ１, ２）を利用する. 本検証実験は, CPU を i5-4250U, メモリを 16GB 用いて実施した.

表 1 の通り, $T_{o2}$ を除き遜色ない処理時間であった. また, 結果の差異も見られなかった.

$T_{o2}$ は処理体系の違いで処理時間に大きな違いが見られた. これは処理体系で用いるファイルの構造に違いがあるため生じたと考えれられる. OPENNESS ではクロールデータから抽出した LINK データを用いて分析を行う. この LINK データは Perl では 1 レコードに複数存在し, システムでは 1 レコードに 1 つのみ存在する. この結果より多くのファイルアクセスが生じたため, $T_{o2}$ の結果になったと考えられる.

しかし, 総合処理時間で考えると処理体系では大きな違いはない. ITHF の提供環境が整備されている時, クロールデータサイズを $S_c$(GB), クロールデータのダウンロード時間を $T_{dc}$, 源データの生成時間を $T_g$, 源データでの分析時間を $T_{ar}$, ITHF のサイズを $S_i$(GB), ITHF ダウンロード時間を $T_{di}$, ITHF での分析時間を $T_{ai}$ とした際,

Perl での総処理時間 $=T_{dc}+T_g+T_{ar}$

システムの総処理時間 $=T_{di}+T_{ai}$

（ただし, いずれもファイルダウンロードを想定すると, ダウンロード速度を $T_{bps}$ とした時, $T_{dc}=S_c/T_{bps}$, $T_{di}=S_i/T_{bps}$ となる. ）

となり, $T_{o2}$ を分析する際の源データ（LINK データ）の生成 $T_g$ はシステムで掛った $T_{o2}$ よりも多くなる.

## 4. まとめ

本報告では, ITHF を使ったシステムの動作検証を行い, 従来の分析手法と比べ ITHF を使った分析手法は遜色ない処理時間が得られることがわかった. 今後は, システムが扱うことの出来る"パッケージ"の充実と実際に ITHF を研究者間に提供するための環境構築が必要となる.

## 参考文献

[1]伊藤, 中平, 三上“国別ドメイン利活用のためのプロビジョンスキーム”情報処理学会第 76 回全国大会（2014）
[2] 難波弘之 “ソーシャルメディアリンク解析に基づいた TLD オープン性評価”長岡技術科学大学 修士課程修士論文（2012）
[3] 三上喜貴“言語間デジタルデバイドの解消を目指した言語天文台の創設”科学研究開発実施終了報告書（2007）
[4] 総務省 情報通信政策研究所 “インターネット検索エンジンの現状と市場規模等に関する調査研究 報告書”（2009）
[5]伊藤, 三上, 中平“クロールデータのプロビジョンスキームにおけるファイル入出力機構の検証”第 13 回情報技術フォーラム(2014)
[6]国立天文台データセンター(2013)「FITS の手引～第 5.3 版～」
[7] 濱 部 勝（2002）「 The IRAF Manual for Beginners」
[8]三上喜貴(2010)「カントリードメインの脆弱性監視と対策」社会技術研究開発事業