D-023

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
Verification of the File I/O Architecture in the Provision Scheme for Crawled Data

伊藤　公　　中平 勝子　　三上 喜貴
Akira Ito　　Katsuko T.Nakahira　　Yoshiki Mikami

1. はじめに

本研究では，研究者間でのクロールデータ統合により研究促進を図るために設計した Information Trade Handle Format(ITHF)[1]を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で円滑に処理出来るか検証を行う．

近年，ソーシャルネットワークサービス(SNS)を提供する Web サイトやそのサイト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日々大量の Web リンクが継続的に生成されている[2]．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量が爆発的に増加しているため，高速処理でそれに対処することが必要となってきた．そこでの課題は，限られたリソースで膨大なデータの処理にかかる時間を始めとする問題である [4].

インターネット上のデジタル・デバイドの分析結果には，分析結果から実態を把握するため確実性が求められる．そのため広い母集団に対してクローリングを行って収集したデータの利用が求められるが，収集するデータには 1 機関のクロール実施環境により数 TB という限界がある．このデータ量は，JP ドメイン上に 2009 年時点に流通しているコンテンツ量(約 7PB[5])から比べると限られたデータ量である．利用するデータを増やすにはクロールを頻繁に実施することが考えられるが，実施のために大量のリソースを確保することが難しいことが往々にしてあるため困難である．その解決策として，統一フォーマットで収集されたデータを流通させる事により利用できるデータが増加し，研究を進める上のコストの減少にも繋がる．

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るため，汎用性の高いファイルフォーマット(ITHF)による大規模クロールデータのためのプロビジョンスキームを提案してきた．ITHF には，インターネット上のデジタル・デバイドの分析に必要な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データ，統計データ等を一つのファイルに統合している．しかし，そのデータ量は現在では数 TB に上るため，提供先の通信環境に合う最適なデータサイズでデータを損なわず提供する事が求められる．そのため，数百 MB 程度のデータ送受信に何時間もかかるような通信環境下で数 TB に上る ITHF をそのまま提供するのは現実的ではなく，統合データ部を分割・圧縮して提供する仕組みを取っている．ITHF の分析処理時には，圧縮したデータを展開して処理するが，元のデータ量が膨大なため分析処理の全工程をメモリ上で処理することが難しく，ディスク上へ一時的にデータを置いて処理をしなければならない[6]．これら処理を含め分析の高速化を目標とするには，分割して提供した ITHF の展開を円滑に行える事が重要である．

本稿では，ITHF を分割する際の最適分割サイズの推定，分割した ITHF を連結する処理を正確に実施できるか検証を行う．

長岡技術科学大学, Nagaoka University of Technology

2. ITHF の利用フローにおける検証項目

クロールデータを用いて ITHF を作成，分析するまでの一連のフローは図 1 の通りである．ITHF はクロールデータや統計データ等をもとに作成するため，そのデータ量は数 TB に及ぶ．ITHF を提供するには，Web インターフェイス等を介してユーザ環境に合う ITHF を提供する事が必要であり，本稿では，2 つの検証項目に対して処理速度という評価基準のもとに検証を行う．

2.1 ユーザの通信環境

ユーザの通信環境によっては，Web インターフェイスを介して以下の選択肢を用意する．

(1) 非分割 ITHF から全データを取得
(2) サンプル調査など取り敢えずデータが欲しい場合を考慮し，取得するデータサイズの指定
(3) 上記でデータ指定がない場合，推奨最適分割数を提示しユーザ側がそれに従うか否かを選択

(3)の選択肢をユーザに提示するのは，日本のように FTTH の通信環境[7]が整っている場合，非分割 ITHF を取得した方が後の連結処理作業が不要となり，処理時間が削減できるためである．しかしながら，このような通信環境が構築できない場合のため，ファイル分割作業を検証し，推奨最適分割数を提示する必要がある．

そこで，1 つ目の検証項目として非分割 ITHF を用いて連結 ITHF をディスク上に展開するまでに要する作業時間から推奨最適分割数の検証を行う．最適な分割サイズで分割 ITHF の作成が行われないと，ランダムアクセス増加によるアクセス速度の低下や分割データの管理が煩雑になるといった問題が考えられる．

2.2 ユーザの分析環境

ユーザの分析環境によっては，非分割 ITHF を分割した後，連結 ITHF を作成するために処理時間が多くかかるという事が想定されるため，連結 ITHF 作成に処理時間がどれ程要するか確認する必要がある．この連結処理の際，分割 ITHF に欠損がないかをチェックしながら，正しく連結 ITHF が作成されるかも確認する．

2 つ目の検証項目として，チェックサムに用いられる様々なハッシュ値を求めるアルゴリズムの処理が，ITHF 処理フロー内で無視できる程であるか検証を行う．連結 ITHF で構築される一連のデータは，メタデータの集合であり連結順序は特に問題とならないが，何らかの原因によって分割 ITHF に破損があるとファイルの連結作業が正しく行われない．この対策として分割 ITHF にハッシュ値を持たせて冗長化を行い，連結処理をする際に求めるハッシュ値と分割 ITHF で保持していたハッシュ値を比較してデータに誤りがないか検出する．

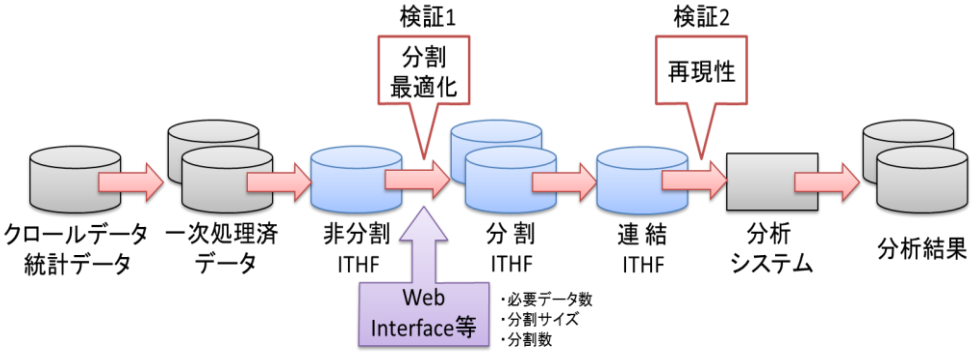

図 1　ITHF 処理フロー

表 1　一次処理済データから連結 ITHF 作成までの処理フローで要した時間

分割サイズ [MB]	非分割 読込時間[ms]	分割 ITHF 書込時間[ms]	分割 ITHF 読込時間[ms]	連結 ITHF 書込時間[ms]	MD5 [ms]	SHA-256[ms]	SHA-512[ms]
100	5,496	1,521	145	44,335	585	1,279	1,114
400	5,328	1,552	120	45,577	580	1,517	1,176
500	5,692	1,493	3,872	48,298	608	1,598	1,202
1,000	5,318	1,478	3,467	45,135	591	1,273	1,120

3. 検証結果

3.1 ITHF 分割

一次処理済データから連結 ITHF 作成までの処理フローで要したディスクアクセスに関わる時間，読込時間と書込時間を分割サイズごとに算出した結果を表 1 に示す．検証実験は，CPU を Xeon E3-1270 V2，メモリを 32GB 用いて実施した．一次処理済データ(約 500MB) 1 行毎に Index 含めて HDU の 1 レコード(2,100 バイト)を作成し，約 13GB のファイルを利用した．

表 1 の通り，分割サイズごとの書込時間に大きな差はなかった．分割サイズによる ITHF 作成時間への影響は少ないものと考えられる．分割サイズごとの読込時間では分割 ITHF を読み込む際，500MB を境に大きく読込時間が変化した．分割 ITHF として提供されるデータは 500MB 未満が読込処理時間を一番削減でき，全体の処理時間削減に有益な分割サイズであると考えられ，この値を基準にして推奨分割数を提示することを提案する．また，処理端末の環境に依っては，ディスクの容量低下等による処理時間の増加が見受けられ，処理時間を平準化する仕組みが求められる．

本稿では，数 GB 単位のクロールデータがある中で比較的少ないデータ量での検証だったが，ディスクアクセスに関わる処理に多くの時間を取られている．今後，より多くのクロールデータが収集されるようになると予想されるため，ディスクアクセスを減らすための工夫が求められる．

3.2 ITHF 連結

ITHF 連結処理における誤り検出アルゴリズムとして MD5，SHA-256，SHA-512 を使い，連結 ITHF 作成時におけるハッシュ値の比較に要する処理時間がどれ程かかるのか検証を行った(表 1)．

どのアルゴリズムにおいても 1ms 程しか処理時間がかからないため，この処理が連結 ITHF の作成時間に与える大きな影響はないものと考えられる．

4. おわりに

本稿では，非分割 ITHF から分割 ITHF を作成する際の適切な分割サイズについて検証を行った．分割 ITHF を読み込むためには，非分割 ITHF を 500MB 未満で分割するのが効率的であると考えられる．しかし，ディスクアクセスに依る時間が ITHF 処理フローの全処理時間の中で大きなボトルネックになることが浮き彫りになり，この影響を最小限とする仕組みが必要である．

現状では閉じた環境で非分割 ITHF を作成しているが，図 1 の Web インターフェイスを実用化するには，外部ユーザが他機関の所有する非分割 ITHF にアクセスしながら分割 ITHF を作成することが求められる．このような仕組みが構築されると，研究ベースで収集されているクロールデータだけではなくグーグルやヤフー等で商用に集められている膨大なクロールデータを一次処理し，流通させるための基盤の一つに繋がる．

謝辞

本研究の一部は学術研究助成基金助成金 24500308 の助成を受けたものである．

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] 中野美由紀,“ビッグデータ統合利活用における課題と技術”,電子情報通信学会誌,Vol.97,No.5 (2014).
[5] 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)
[6] 定兼邦彦, “ビッグデータのための簡潔データ構造”,電子情報通信学会誌,Vol.97,No.5 (2014)
[7] 総務省, “平成 25 年度版情報通信白書” (2014)