4L-6

国別ドメイン利活用分析のためのプロビジョンスキーム

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，収集した大規模 Web コンテンツの提供から分析までを行うプロビジョンスキームの設計を行う．

近年，インターネットの普及は大量のネットユーザとその利用行動履歴データ(一種のビッグデータ)を生み出した．更にコンピュータの高速化がそのビッグデータの即時処理，データ分析を容易にし，ビッグデータを分析する試みが活発になっている．様々なビッグデータの円滑な利用に向けて ITU[1]や ISO/IEC ワーキンググループ[2][3]を始めビッグデータを扱う機関ではその標準化の議論を始めている．ビッグデータの種別に則したデータ提供法の標準化が行われることで，大容量データの高速処理・効率化が期待され学術的・経済的な利益を実現できる．

ビッグデータの標準化により利益を受ける分野の 1 つに，Web コンテンツによるデータマイニング分野が考えられる．研究としては，デジタルデバイドの実態調査としてこれまで行われてきたリンク活用[6]や言語利用[7](本稿では，利活用と定義)を用いて Web コンテンツの利活用を分析する場合が挙げられる．国別ドメインごとに利活用を分析することで，国内や国家間での Web コンテンツへのアクセス性や言語多様性などの Web コンテンツに着目したデジタルデバイドが観測可能になる．しかし，現在の Web ページ総数は 1 兆ページを優に超え，1 機関でその全ての収集・分析は不可能である．そのため，同じ目的意識を持つ集団で Web コンテンツのプロビジョンスキームを構築すると各機関の Web コンテンツを有効利用でき，分析の効率化や分析精度の向上，更に様々なデータを組み合わせることで新たなイノベーションが生まれる可能性がある[8]．

本稿では，そのためのプロビジョンスキームの設計を行い，その実現に必要なデータ提供フォーマットを提案する．

2. プロビジョンスキーム

データ観測機関間のプロビジョンスキームに成功した分野の 1 つが天文学である．その理由は幾つか存在するが，技術面という観点から考えるとデータセンターの存在とデータを記録するフォーマットの存在である．天体観測で得られる撮像/分光データは，観測所ごとに生成され，専有期間が過ぎて二次利用可能となったものはデータセンターに蓄積され，広く公開されている．データを公開しても問題なくデータの利用が出来るのは，データを記録するための標準フォーマット，FITS[4]の存在がある．FITS が策定されたことで，データセンターで問題なくデータの公開が出来るようになり，FITS のデータ解析ツールがいくつも開発されている(例えば，IRAF[5])．

この仕組みを Web コンテンツに適用するとほぼ同一のプロビジョンスキームの構築が可能になり，様々な研究領域で応用できるはずである．図 1 に Web コンテンツのプロビジョンスキームのこれまでと今回提案する手法を示す．提案スキームでは，「(1)データ保有，(2)データ分析システム，(3)分析結果の共有，(4)(1)の情報を集めたデータベース」の 4 ステップからなる．提供データは，収集された Web コンテンツを一定の法則で一次処理したものが各機関にて保有される．各機関におけるデータ保有情報は専用のデータベースに保存され，必要に応じて研究者に提供可能とする．提供されたデータは，独自の分析システムで分析される．こうして分析された結果は論文として公開され，その情報は研究者に共有される．この繰り返しにより国別ドメイン利活用分析のためのプロビジョンスキームは円滑に作用すると考えられる．

データ提供のため一次処理したデータの標準規格として ITHF(Information Trade Handle Format)を設計した．分析システムは，大きく分けて 3 つの機能を実装する．これら機能の他に，ITHF ヘッダを更新する際の書き込み機能を実装する．各分析はパッケージ単位での実施に対応し，様々な分析に対応する仕組みを構築する．

【IHTF の入力】格納データを正確に読み取る．

【分析パッケージの入力】出力したい分析結果に合致する分析パッケージを選択する．

【分析結果の出力】分析パッケージを使って ITHF を分析し，結果を出力する．

The Provision Scheme for Analysis of the ccTLD Utilization
Akira Ito, Katsuko Nakahira, Yoshiki Mikami
Nagaoka University of Technology

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.

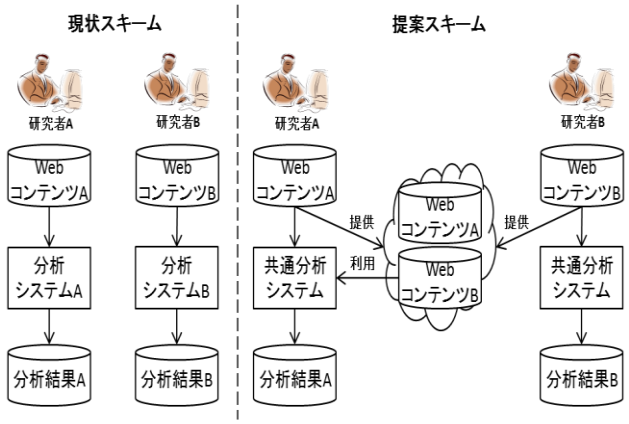


Fig.1 提案するスキーム

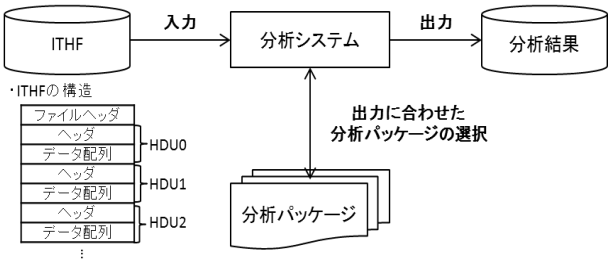


Fig.2 分析システムの概要

Table.1 非圧縮と圧縮時の比較

	格納データ(Ascii)	非圧縮	圧縮
ボリビア	1,570MB	6,085MB	54MB
パラグアイ	1,066MB	2,402MB	22MB
対象国平均	281MB	1,095MB	9MB

3. ITHF フォーマット

ITHF はファイル形式を識別するための情報(ファイルヘッダ), ヘッダとデータ配列の構造(HDU, Header Data Unit)をしている. HDU0 には公的機関が出している統計・調査データを, HDU1 にはクローラー等で収集した Web コンテンツ中の URL データを, HDU2 以降にはこれまでの収集データを 1 次処理したデータを格納する(例えば, Web ページの言語解析結果). HDU のヘッダは, "キーワード＝値/コメント"を 1 レコードとする構造をしている. キーワード(以下, KW)には, 次の 3 種類が定義される. KW は HDU 自身に関するもの(作成日, 観測対象国等)や Web データ取得に関するもの(収集日時, 収集に使用したソフトウェア名等), データ格納に関するもの(データの次元数, フィールド長等)等が使用される.

【必須 KW】必ず記載する KW

【重複 KW】統計・調査データの記載時, 年ごとに重複して生成される KW

【予約 KW】必須ではないが, 予約された KW

格納データは大容量であり提供・分析に不向きである. 特にインターネットが発展している国は大容量の Web コンテンツが収集されるため(例えばブラジル, リンクのみで約 10GB), 一般的に容量の削減, 処理の高速化に繋がる Binary, gzip 圧縮形式を採用した. 表 1 に示す通り, このプロセスを経ることで ITHF の容量は非圧縮提供に比べ, 約 85%削減される. 提供に適する形となり, 分析も処理の工夫により今までより高速化が見込める.

4. まとめ

本報告では, 大規模 Web コンテンツを効率的に分析・提供するため, 従来スキームとは異なるプロビジョンスキームを提案した. このプロビジョンスキームでは ITHF について設計を行い, 提供に適するか検証を行った. この ITHF を用いることで, 各研究者が独自に収集・分析を実施するより効率的に分析が出来るスキームを提案した. また, このプロビジョンスキームに則った分析システムの概要も提案した. 今後は分析システムの設計・構築を行い, 従来システムと比較して優位であるかを評価したい.

謝辞

本研究の一部は科研費 MEXT/JSPS(24500308, 代表:長岡技術科学大学 中平勝子) の助成を受けたものである.

参考文献

[1]ITU:Big Data: Big today, normal tomorrow, ITU-T Technology Watch Report (2013)

[2]Gim,Lim,Jung: Next Generation Analytics & Big Data (A Reference Model for Big Data), ISO/IEC JTC1 SC32 Ad-hoc meeting (2013)

[3]Klaus-Peter Eckert: Big Data and Next Generation Analytics, 28th ISO/IEC JTC1 Plenary Meeting (2013)

[4] 国立天文台 天文データセンター: FITS の手引～第 5.3 版～ (2013)

[5] 濱 部 勝 :The IRAF Manual for Beginners (2002)

[6]三上喜貴: カントリードメインの脆弱性監視と対策, 科学研究開発実施終了報告書 (2011)

[7]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)

[8]伊藤,福原,他: 大規模かつ多様なデータをリアルタイム解析のための情報基盤, 情報処理学会 全国大会講演論文集, pp.373-375 (2013)

Copyright ©2014 Information Processing Society of Japan.
All Rights Reserved.