情報格差観測のための分析システムの開発

新井 裕樹　中平 勝子　三上 喜貴

長岡技術科学大学 工学研究科

国別の情報格差の実態を評価する指標として，ネットワーク環境の普及率，パソコン台数等の情報基盤に関する格差指標が取り上げられることが多い．しかし，情報格差の実態を捉えるためには，実際の利用状況やそれが生み出す効用の程度まで評価することが必要であり，このような認識にたって，筆者らは情報格差を評価するための指標開発とこれに必要なデータの収集と解析を行うシステムの開発を行ってきた．本稿では，このための基盤／利用／効用の三階層からなる指標体系の全体構想，および現在開発中の情報格差分析システム"*i*-GALAKSY"の概要について報告する．

An analysis system for global digital divide observation

Yuki Arai　Katsuko T. Nakahira　Mikami Yoshiki

Nagaoka University of Technology

When the digital-divide between countries is discussed, those indicators which represent the level of ICT infrastructure, such as telecommunication network penetration ratio and the number of personal computers, are often used to evaluate the degree of the divide. But more realistic view of the divide can only be illustrated by a set of indicators which represent not only the level of infrastructure but also the level of actual usage of, and the benefits provided by those infrastructure. Authors are developing such a set of indicators and are developing an integrated system to collect and compile those set of indicators. This paper presents an outline of the three-tier (infrastructure/usage/utility) set of indicators and the system.

1　はじめに

情報社会と呼ばれる現在，人はインターネットを介して世界中のあらゆる情報へ瞬時にアクセスできる．しかし，その恩恵を享受することの可能な人と不可能な人との間に生じる情報格差（デジタル・ディバイド）は解消されるどころか，むしろ拡大する傾向にすらある．木村忠正は『情報ネットワークへのアクセスを「もつ」「もたない」が，社会階層（国家）により大きく異なり，しかも，そうした階層間（国家間）の経済的格差，社会的格差が拡大する傾向にある』と述べている [1].

特に国家間の情報格差については国連のミレニアム開発目標でもその解消に向けて目標が設定され，また，ITU などの国際機関では，目標の達成状況をフォローするために，Digital Access Index[2] をはじめとして，ネットワーク環境，パソコン台数等の一次データから構成される様々な情報格差指標が開発されてきた．しかし，既存の指標の多くは，通信基盤の投資度合や充実度のみを考慮したものであり，実際の利用実態に関する格差を考慮に入れたものは少ない．国連のミレニアム開発目標も人口 100 人あたりの電話回線数，携帯電話契約数及びインターネット利用者数という三指標を取り上げているに過ぎない．情報格差が情報ネットワークへのアクセス機会の格差に起因することを考えれば，通信設備やパソコンの台数などの基盤を主体に格差を計測し，評価することは格差観測にあたっての必要条件ではあるが，実際の情報ネットワーク利用状況やそれが生み出す社会・文化的価値である効用といった段階でも格差が生じていると予想されることから，真の情報格差の実態を捉えるためには，「基盤」「利用」「効用」といった重層的な構造を持つ指標群を開発する必要がある．また，実際に調査した結果を公表し，情報格差に関する政策提言の参考情報として使用することが指標開発の目的であるため，情報格差

の現状を公表する仕組みづくりも必要となる．筆者らはこうした指標・システムを開発するに当たって，言語天文台，サーバ巡回クロウラー技術，ccTLD の利用実態分析，情報格差指標，リンク構造解析の 5 つの研究を参考とした．

これらの研究は，いずれも情報格差の観測を目的としたものであり，また，そのために独自のソフトウェア・ツールを開発している．しかし，これらの研究は一定の問題意識を共有しつつも，観測のために開発されたソフトウェアツールはシステムとして統合されていない上に，継続的に計測できるような構成となっていない．このため，情報格差の指標を継続的に常時観測できるツールとして統合する必要がある．さらに，我々が開発した指標だけでは情報格差指標としては不十分であるため，指標を管理することのできる機能も必要である．本稿では，これらの情報格差指標に基づいて調査される統計データを系統的に分析するためのシステム，"*i*-GALAKSY（internet-Governance And LAnguage Knowledge-base SYstem）"を開発した．

2 先行研究

指標・統合システムを開発するに当たり，次の 5 つの研究を参考とした．

三上ら [3] の言語天文台プロジェクトは，言語間デジタルデバイドの解消を目指して，インターネット上の言語活動を観測する言語天文台を構築し，言語間デジタルデバイドの計測を行っている．インターネット上の言語活動を把握することで母語や，使用文字コードの現状について把握し，少数言語の参加を促す手段を見つけ出す活動を行っている．イタリアのミラノ大学が開発した UbiCrawler[4] と独自開発の言語判定エンジン LIM[5] を用いている．

中平ら [6][7][8] はアジア・アフリカドメインにおけるネットワーク環境に関する調査研究の一環として，ネットワークの物理的基盤であるサーバの設置状況を調査し，アフリカやアジアでは，国内設置の割合が非常に低いことを明らかにした．また，通信パフォーマンスの観点からもアフリカの情報基盤格差は深刻なものと報告している．さらに，こうした実態を継続的に，かつ相手国の通信回線への負荷をなるべくかけずに調査するためのサーバ巡回クローラを開発した．

和田ら [9][10] は島嶼国 ccTLD の利用実態に関する調査研究の一環として，割り当てられた ccTLD について，当該地域のユーザが実際に利用しているかどうかを調査し，島嶼国・地域の ccTLD の多くが住民一人当たりページ数が世界平均よりも多くのは，主として海外向けのウェブホスティングサービスが行われていることによるものであることを示唆している．この研究は，これらの島嶼地域では空ページやリダイレクションページなどを使用した Web スパムページとして利用されている可能性が高いことを報告している．

上嶋ら [11] はデジタルデバイドを基盤，能力，効用の 3 階層に分け，基盤以外の指標を多く開発している．また，国家間，地域間のデジタルデバイドの不平等度の指標化としてジニ係数の考え方を適用し，デジタルデバイドの分析に活用し得ることを報告している．

石原 [12][13][14] は海外報道機関へのリンク率の分析を行った．国毎の著名な海外報道機関へのリンク数を調査することで，インターネットに自由にアクセスできる国において，既存メディアが十分に役割を果たせていない場合はインターネットが海外ニュースにアクセスする代替メディアとして使用されているという実態を報告している．

また，Web 空間の発達度を評価するためにグラフ理論を応用し，outdegree 分布，ドメイン内における強連結成分の相対的大きさと平均距離，他ドメインへのリンク数と平均距離，到達可能性から構成される Web 環境の自由度，ドメイン内での結合度，ドメイン外への結合度といった 3 種類の指標群からなるメトリクス体系を提案し，その有効性を報告している．

3 情報格差

3.1 情報格差定義

まず，本稿における情報格差の定義を行う．情報格差は物理的な通信手段の有無のみをもって測ることはできない．情報通信技術を扱う能力や情報通信技術によって生み出される効用の程度まで評価する必要があるために，情報格差を表 1 に示すように 3 段階で定義した．

第一階層の「基盤」は，ネットワークの物理回線や PC 端末などの物理的な基盤を指す．第二階層の「利用」は，情報基盤技術を利用し，インターネット上で取り扱われているデータ・情報の量である．第三階層の「効用」は，情報技術を利用したことによって生み出されたインターネット上の経済・社会・文化的価値である．

表 1　情報格差の 3 階層

効用	情報技術の利用によって生み出される価値
利用	情報基盤技術を利用したデータ，情報の量
基盤	情報基盤技術に対する投資の充実

3.2　情報格差指標

次に開発するツールで取り扱う情報格差指標について説明する．情報化指標は，ITU などの外部機関によって開発・利用されている外部データと，我々がクローラを用いて収集したデータから測定可能な新開発の指標の二通りに分けられる．

外部データを利用する指標としては，固定電話回線数，携帯電話契約者数，PC 台数，インターネット利用者率，ドメイン数などが考えられる．これらの指標は ITU や ISC[15] などの公的機関のデータを使用するものである．

独自データに関する「基盤」の指標としては通信パフォーマンス・Web サーバ数・Web サーバ国内設置比率，「利用」の指標としてはページ数・文字コード利用率・クッキー利用率・空ページ比率・リダイレクションページ比率，「効用」の指標としては母国語ページ比率・海外報道機関へのリンク率・Web 環境の自由度・ドメイン内での結合度・ドメイン外への結合度を開発した．これらの指標の意味・測定方法を表 2 に付録として示す．

次に，独自データに関して Web サーバ国内設置比率，リダイレクションページ率，海外報道機関へのリンク率の指標を取り上げ説明する．Web サーバ国内設置比率は，当該 ccTLD の下にある Web サーバのうち国内に設置されているものの割合を示す．通信基盤が整備されていない国では国外にサーバを設置するため，当該 ccTLD に対応する国内にサーバが設置されない現象が発生する．そのため，この指標は ccTLD ごとの通信基盤の整備度合いの格差を表す．リダイレクションページ率は，スパムページにリダイレクション情報が記載されていることからスパムページの存在を示唆する指標である．そのため，ccTLD 別のリダイレクションページ率は悪質な Web ページの利用の状況の差を表す．海外報道機関へのリンク率は，ccTLD 毎の著名な海外報道機関へのリンク数を調査することで，インターネットに自由にアクセスできる国において，既存メディアが十分に役割を果たせない場合はインターネットが海外ニュースにアクセスする代替メディアとして使用されているという実態を表す．そのため，ccTLD 別の情報規制の差から生まれる社会的格差の現状を表す．

4　システム

i-GALAKSY の概要を図 1 に示す．*i*-GALAKSY は，大きく入力部・出力部，およびそのインターフェース・データベースから構成される．また，恒久的運用を意識したため，特に管理部を設けている．以下，*i*-GALAKSY を構成する各部を概説する．

4.1　インタフェース

インタフェースは，Web サーバ上で動作する Web アプリケーションとコマンドラインを実装する．Web ブラウザからはユーザがシステムに対してデータの入力・情報管理・出力に関する要求を出すことが可能である．入力は，軽量なデータをファイルとして入力できる．情報管理は，クローラに対する制御情報を与えることや指標のマネジメントを行うことができる．出力では，情報格差指標の測定とユーザの要求したクエリの結果を取得できる．データベースの処理結果を HTML 画面，CSV ファイル，グラフ表示，MapServer[16] でのマッピングなどの形式で出力するかを要求できる．コマンドラインからはユーザがシステムに対して，大容量のクロウリングデータを入力する際に使用する．

4.2　入力部

入力部は収集フェーズと抽出フェーズから構成されている．以下，それぞれのフェーズの説明をする．

収集フェーズは，インタフェースで入力された情報・制御情報を元に，ファイルから情報を収集するか，インターネット上から Web クローラを使ってインターネット上から情報を収集するフェーズである．クローラはミラノ大学開発の UbiCrawler と独自開発のサーバ巡回クローラを使用する．UbiCrawler は Web ページを収集することを目的としたクローラ，サーバ巡回クローラはサーバの情報（サーバ

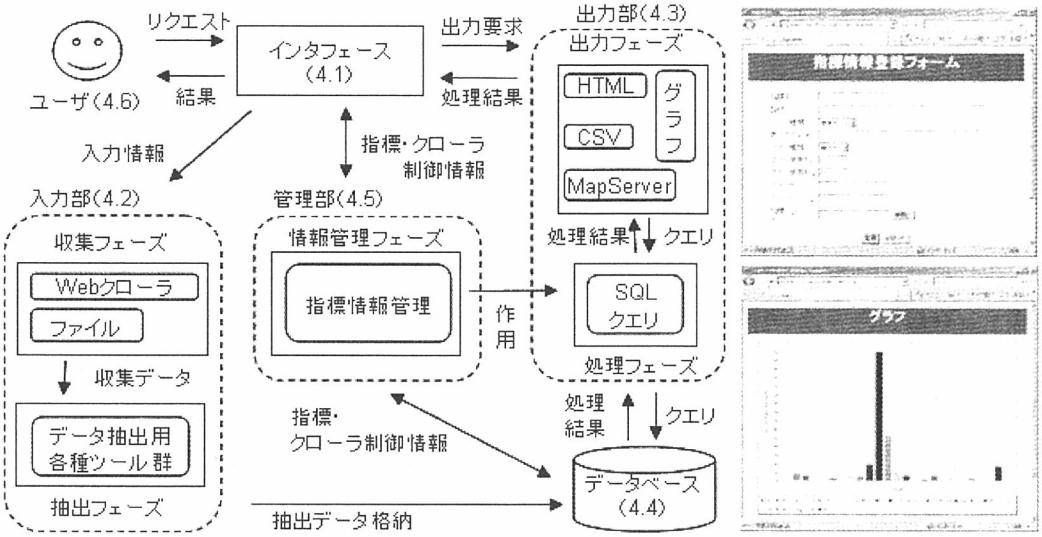


図 1　システム概要図

の IP アドレス，サーバ所在国など）を取得することを目的としたクローラである．

データは，集めただけでは扱うことができないためデータから情報格差観測に必要なものだけを抽出する必要がある．そのため，抽出フェーズでは Web クローラが収集したデータから各種データベースに格納する必要な情報を各種ツールを使用し，抽出する．

ツールは，以下のように Ubi-API[4]，HTML パーサ，GeoIP[17]，LIM，リンク構造分析ツールを使用する．Ubi-API は，UbiCrawler が収集したデータから HTML ページコンテンツや HTTP ヘッダ情報を取得するためのものである．HTML パーサは，HTML ページの META タグに記載されているリダイレクション情報等を抽出するために使用する．GeoIP は，IP アドレスからサーバの地理情報を求めるためのツールである．LIM（Language Identification Module）は，ウェブページやテキストデータの使用言語，使用文字体系，使用エンコーディング方式を自動的に判定できる言語判定エンジンである．LIM は，N-gram を利用して言語判定を行っており，現在約 290 言語を識別できる．リンク構造分析ツールは，グラフ理論を用いて Web 環境の自由度，ドメイン内での結合度，ドメイン外への結合度，海外報道機関へのリンク率などのリンク構造を ccTLD 別に分析するための独自開発のツールである．

これらのツールを必要な情報を取得する際に使用し，その結果をデータベースに渡す．

4.3　出力部

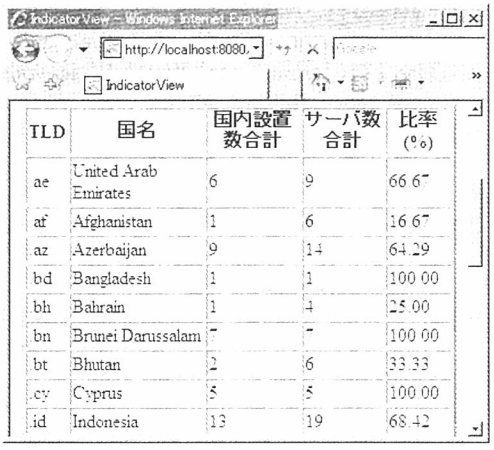
出力部は処理フェーズと出力フェーズから構成されている．以下，それぞれのフェーズの説明をする．

処理フェーズは，要求のあった SQL クエリを実行し，処理結果を返すフェーズである．情報格差指標を計測するための SQL クエリは，データベースに蓄積する構成となっている．そのため，情報格差指標の出力が要求されるたびにデータベースに格納されている SQL クエリ情報を元に指標の計測を行う．

出力フェーズは，ユーザから要求のあった情報格差指標または，クエリの結果を HTML，CSV，グラフ，MapServer の四つのいずれかの方法を用いて出力する．MapServer は，オープンソースの地図描画エンジンである．現在実装されているものは，通信パフォーマンスごとに各国を色分けし，通信速度の差を世界地図上で閲覧することができるものである．図 2 は Web サーバ国内設置比率を HTML 上で表現した現在開発中の画面例である．

4.4　データベース

このシステムでは，大量の Web ページの情報を扱う．実際，現在までに収集を行った Web



TLD	国名	国内設置数合計	サーバ数合計	比率(%)
ae	United Arab Emirates	6	9	66.67
af	Afghanistan	1	6	16.67
az	Azerbaijan	9	14	64.29
bd	Bangladesh	1	1	100.00
bh	Bahrain	1	4	25.00
bn	Brunei Darussalam	7	7	100.00
bt	Bhutan	2	6	33.33
cy	Cyprus	5	5	100.00
id	Indonesia	13	19	68.42

図 2　Web サーバ国内設置比率画面例

ページは 6 千万ページを超えている．今後も継続的にページを収集を行うことを考慮すると，単一クエリーの実行時間が遅くなることが考えられる．そのため，単一クエリーの実行時間を高速化するためにデータベースを分散化する．現在データベースは PostgreSQL を使っており，PostgreSQL を分散化するため手段としてミドルウェアの pgpool-II を採用した [18]．単一クエリーの高速化は pgpool-II のパラレルクエリー機能を使用することによって可能である．パラレルクエリー機能は，論理的に一つの表を物理的に異なるホストに分割・分散し，ホストごと並列にクエリーをそれぞれ実行することで単一クエリーの実行時間の短縮する．

4.5　管理部

管理部は情報管理フェーズから構成されている．以下，情報管理フェーズの指標情報管理の機能を説明をする．

指標情報管理は，Web 上から指標情報（指標名，指標計測用の SQL など）を追加，変更，削除することのできる管理機能である．本システムの提供する情報格差指標は，政策提言の参考情報や政策効果のフォローアップ指標として使用することが目的であるため，継続的にシステムを運用していくことが必要である．この過程で，指標の算出方法や指標体系の構成に変更が生じることもありうる．その際には原データを保持しておき，新しい指標算出方法に基づいて再計算するといったことにも対応する必要がある．本システムはこうした要求に対応するため，UbiCrawler 及びサーバ巡回クロウラーという二つのクロウラーの収集した原データをそのまま保持する設計となっているほか，指標の算出方法も，プログラムを変更することなくユーザが新しい SQL クエリーの形で計算式を再定義するだけで対応できるようになっている．

4.6　利用ユーザ

このシステムを利用するユーザは，一般ユーザと管理ユーザの 2 種類に分けることが必要だ．システムの管理部分に必要不可欠にアクセスされるのはセキュリティ上好ましくないためだ．そこで，一般ユーザと管理ユーザについて説明する．一般ユーザは各情報格差指標の計測結果を取得する人を想定し，視覚的に情報格差指標を把握することを求める人，実際に情報格差指標の数値情報を把握・取得することを求める人が考えられる．視覚的に情報格差指標を把握することを求める人は，視覚的に情報格差が把握可能な MapServer，グラフが機能し，実際に情報格差指標の数値情報を把握・取得することを求める人は，数値として情報を与える HTML，CSV 出力が機能する．このため，一般ユーザがアクセス可能な機能は出力部だけである．

管理ユーザは，データの入力・収集，クローラ・指標の管理，データの出力，データベース接続など全機能にアクセスする人を想定する．そのため，入力部，管理部，出力部の全機能をアクセス可能とする．まだ未実装であるが，ユーザ ID とパスワードによるユーザ認証機能を今後実装し，一般ユーザと管理ユーザを区別し，管理する予定である．

5　運用

本システムは継続的に動作し，情報格差指標を提供するため，システム運用を行う上で課題がいくつか生じる．ここでは，システム運用についての説明を行う．

システムを持続的に運用する上で発生する問題として，人的要因と技術的要因が考えられる．人的要因は，システムを管理する容易さをどのように実現するかが問題である．現時点では，データベースシステムのオンラインリカバリツール，サーバのモニタリングツールやクローラ制御情報管理ツール（クローリングタスクの設定や，動作状況閲覧）を Web で提供し，

システムに対するアクセシビリティを向上することで，運用の容易さを実現することを考えている．

技術的要因の問題は，継続的にシステムを使う上でどのように安定的にデータを提供するか，システムのデータ量が増加することに対してどのように対処するかの二つが考えられる．継続的にシステムを使う上でどのように安定的にデータを提供するためには，物理的な面から考えるとハードディスクを RAID に対応することで，突発的な障害に対応することが望ましい．論理的な面から考えると，データベースの分散化を行い，レプリケーションを適応することでデータの破損などの障害に対応し，データを安定的に提供することが挙げられる．また，継続的にシステムを運用することでデータ量が増加することに対しては，論理的に 1 つの表を物理的に異なるデータベースに分散化し，データの削減を行うことで対処することが考えられる．

6 まとめ

本論文では，情報格差観測のための指標と情報格差指標を取り扱うシステム”*i*-GALAKSY”について述べた．今後は，データベースの分散化，ユーザ認証機能を検証・実装化し，情報格差を観測していきたい．

参考文献

[1] 木村忠正：デジタルデバイドとは何か，岩波書店（2001），p.10.

[2] International Telecommunication Union, http://www.itu.int/net/home/index.aspx

[3] 児玉茂昭，チュー ユー チョーン，三上喜貴：自動言語判定手法の開発とそれを利用したインターネット上の言語分布に関する調査，日本言語学会第 135 回大会，松本，2007.

[4] Ubicrawler, http://law.dsi.unimi.it/.

[5] 中鉢欣秀，Gondri Nagy Janos 他：言語天文台を設立するための言語判定フレームワークの開発，第 171 回自然言語処理研究会，2006.

[6] Katsuko T. Nakahira et. al : Geographic Location of Web Servers under African Domains, The 15th International World Wide Web Conference, Edinburgh, 2006.

[7] Katsuko T. Nakahira et. al : Low-Load Server Crawler: Design and Evaluation, The 17th International World Wide Web Conference, Beijing, 2008.

[8] 星野哲哉，中平勝子，三上喜貴：サーバ接続環境調査のための低負荷クローリング手法の検証，FIT2007，pp83-86，2007.

[9] 和田祥太：島嶼国 ccTLD の有効活用と管理改善のための提言，長岡技術科学大学大学院工学研究科修士論文

[10] 和田祥太，中平勝子，三上喜貴：島嶼国 ccTLD の利用実態，電子情報通信学会信越支部大会，pp152，2006

[11] 上嶋智大，中平勝子，三上喜貴：デジタルデバイドの評価指標についての一提案，FIT2007，pp481-484，2007.

[12] 石原直幸：グラフ理論を用いたカントリードメインのリンク構造解析，長岡技術科学大学大学院工学研究科修士論文

[13] 石原直幸：グローバルニュースメディアとしてのインターネット，政策空間 http://www.policyspace.com/,vol.44,2007 年 6 月．

[14] 石原直幸，中平勝子，三上喜貴：ccTLD を単位とした Web コミュニティ構造の分析，FIT2007，pp115-118，2007.

[15] Internet System Consortium, http://www.isc.org/.

[16] University of Minnesota, MapServer, http://mapserver.gis.umn.edu/.

[17] MaxMind co. ltd., GeoIP, http://www.maxmind.com/.

[18] pgpool-II, http://pgpool.projects.postgresql.org/pgpool-ja.html

表 2　付録：情報格差指標

	指標名	説明	測定方法
基盤	通信パフォーマンス	ccTLD 別の通信基盤の整備状況を観測基点からの応答速度として示す.	サーバ巡回クローラで Web サーバの反応時間を ccTLD 別に計測.
	Web サーバ数	コンテンツ利用のための前提となるサーバに対する ccTLD 別の投資度合を測る.	サーバ巡回クローラで取得した Web サーバ数を ccTLD 別に計測.
	Web サーバ国内設置比率	当該 ccTLD の下にある Web サーバのうち国内に設置されているものの割合. 国別の国内の通信基盤に対する投資度合，通信環境の整備度合を示す.	サーバ巡回クローラでサーバのドメイン情報，IP アドレスを取得し，GeoIP で IP アドレスに対応するサーバ設置国を調査.
利用	ページ数	ccTLD 別の利用ページ数を示したもので，情報技術を利用する度合を示す.	UbiCrawler で取得したページ数を ccTLD 別に測定.
	母国語における異種文字コード利用数	母国語表記のために用いられている文字コード数で，文字コードの混乱状況を表す.	UbiCrawler で取得した情報から LIM により言語判定を行い，使用文字コードを ccTLD 別にカウント.
	クッキー利用率	ccTLD 別の EC サイトの利用率を表す指標，特にショッピングサイトなどで利用されることから E コマースの利用状況を間接的に表す.	UbiCrawler で取得した情報から，Ubi-API を用いてクッキー情報を抽出し，クッキーの有無を ccTLD 別に測定.
	リダイレクションページ比率	リダイレクションに関する情報を記述しているページの比率. リンクスパムの可能性を示唆している.	UbiCrawler で取得した情報から HTML パーサを用いて refresh 情報を抽出し，リダイレクションページ数を計測.
	空ページ比率	コンテンツに何も記述されていないページの比率. リダイレクション用のページにはコンテンツを記述する意味は薄いため，コンテンツがゼロであるようなページはリダイレクションなどを目的とした Web スパムの可能性を示唆している.	UbiCrawler で取得した情報から空ページを取得し，ccTLD 別に測定.

	指標名	説明	測定方法
効	母国語ページ比率	当該 ccTLD における母国語ページの比率を表し，文化的な価値を生み出しているかどうかを示す．Web 上で母国語の表記可能か否かにより，母国語による情報伝達の可否が決定する．そのため，母国語による情報伝達手段が存在しない場合は，Web 上での情報取得・公表に弊害をきたすため，言語間における格差が生じる．	UbiCrawler で取得した情報から LIM でページの言語を判定して母国語ページ数の比率を計測．
用	Web 環境の自由度	ページの作成やリンク形成が何ら制約を受けずに行われる場合 outdegree 分布かべき乗側に従う，という規則に基づき Web 環境の自由度を表す．	outdegree 分布のべき乗則適応度によって計測．
	ドメイン内での結合度	当該 ccTLD 内でリンクをたどることで到達可能なページがいかに多くあり，互いに強く結ばれているのかを示す．	当該 ccTLD 内の各ページの強連結成分（SCC）のサイズ，及び SCC 内の頂点間平均距離を求める．
	ドメイン外への結合度	他の ccTLD へ短距離で到達することが可能かを示す．ドメイン間の情報流通についての傾向をつかむことができる．	他ドメインへのリンク比率と近接ドメインへの平均距離，到達可能性を求める．
	海外報道機関へのリンク率	インターネットが既存メディアに対する代替メディアとして使用されているかどうかを示す．	UbiCrawler で取得したページからリンクを抽出し，ccTLD 毎に著名な海外報道機関へのリンク数を集計．