計量書誌学の新たな挑戦
－国産オルトメトリクス計測サービスの開発－

吉田　光男*

インパクトファクターや h 指数など被引用数をもとにした指標により，学術情報の評価が行われてきた。2010 年前後より，被引用数にかわる評価指標として，ソーシャルメディアなどにおける言及の利用が試みられている。この指標はオルトメトリクスと総称される。本稿では，オルトメトリクスの概要，および，筆者が開発している国産オルトメトリクス計測サービス Ceek.jp Altmetrics の開発について述べる。さらに，ソーシャルメディアなどにおける日本語文献の言及動向を報告し，今後の展望について述べる。

キーワード：学術情報，オルトメトリクス，ソーシャルメディア，クローラ，ウェブ API

1. はじめに

50 年以上も前から，学術研究の定量評価が試みられており，学術雑誌に関してはインパクトファクター（Impact Factor）[1]が，研究者に関しては h 指数（*h*–index）[2]が有名である。これらの指標は学術文献の被引用数をもとにしている。被引用数は，引用されることは学術的に意味があることを仮定しており，プロ（研究者）が評価を行うという点において，比較的信頼できる。そのため，文献そのものの評価指標としても被引用数が使われる場合が多い。しかし，ある学術文献が引用され始めるまでに，出版されてから 2 年から 3 年の時間を必要とする[3]。つまり，ある研究が社会的に注目されていたとしても，被引用数を使う限り，その注目を定量評価できるのはずいぶんと後のことである。

2010 年前後より，被引用数にかわる評価指標としてオルトメトリクス（Altmetrics）が注目されている。オルトメトリクスは「alternative metrics」から作られた造語であり，文献の閲覧数，ブログやソーシャルメディアでの言及，マスメディアでの報道など，社会的な影響を加味した文献評価指標である[4,5]。この指標には，引用文献の出版を待たず，早期に影響度を計測できるという利点がある。Altmetric.com[6]や Impactstory[7]といったオルトメトリクスを計測する商用サービスも提供され始めている。

本稿では，計量書誌学の新たな挑戦としてのオルトメトリクスに着目し，筆者が開発している国産オルトメトリクス計測サービス Ceek.jp Altmetrics[8]について述べる。まず，オルトメトリクスについて解説し，次に筆者が開発しているサービスについて述べる。そして，それらを踏まえた今後の展望について述べる。

＊よしだ　みつお　豊橋技術科学大学 情報・知能工学系
〒441-8580　愛知県豊橋市天伯町雲雀ヶ丘 1-1
（原稿受領　2014.11.4）

2. オルトメトリクス

オルトメトリクス（Altmetrics）は「alternative metrics」から作られた造語であり，学術雑誌に関する評価指標であるインパクトファクターや，研究者に関する評価指標である h 指数を補完する新たな指標として，2010 年前後より注目されている。インパクトファクターや h 指数は学術文献の被引用数をもとにしており，プロ（研究者）による評価である一方で，いわば同業者による内輪な評価にとどまっているという問題がある。オルトメトリクスはそれにとどまらず，文献の閲覧数，ブログを含むソーシャルメディアでの言及，マスメディアでの報道など，研究者以外の関与を含めた社会的な影響を示す様々な視点を組み入れることにより，文献が社会に及ぼした影響度を包括的かつ早期に計測することを目指す指標である[4,5]。

オルトメトリクスは，通常，学術文献における影響は加味されておらず，あくまでも，文献が人々に及ぼした影響や，人々が示した興味および関心を示すものであることに注意する必要がある[9]。また，人々の興味および関心には，ポジティブな感情とネガティブな感情があることに留意する必要がある。つまり，単純に言及数を集計している現状では，言及数が多いからといってポジティブな影響を及ぼした文献であるとは限らない。むしろ，ネガティブな感情を呼び込むような文献の方が，より多くの興味関心を引くという調査結果も存在する[10]。もちろん，このような言及意図を加味しない問題は，被引用数をもとにしたインパクトファクターや h 指数にも同様に存在する。

オルトメトリクスは注目され始めてから 5 年程度が経過しており，Altmetric.com や Impactstory といったオルトメトリクスを計測する商用サービスも提供され始めている。また，出版文献の影響を可視化するために，出版者自身がオルトメトリクスを計測する試みも行われている。

Altmetric.com[6]は Euan Adie らによって 2011 年から準備され，2012 年 2 月より開始されたオルトメトリクス計測サービスである。一般の利用者には，図 1 のようなドーナツ型のバナーによって，各文献のオルトメトリクスが提

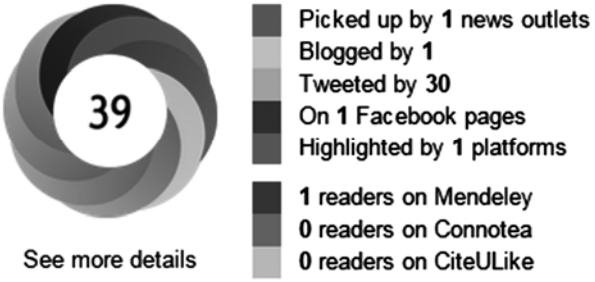


図 1　Altmetric.com が提供するドーナツ型のバナー

供されているサービスとして知られている。出版者や学術機関向けには，より詳細な言及状況や，学術雑誌単位の影響度などの情報が有償で提供されている。Altmetric.com は主に DOI（デジタルオブジェクト識別子）が付与された学術文献を計測対象としており，日本語の文献の大半には DOI が付与されていない現在，計測対象となる日本語の文献はごく少数であると推察される。

Impactstory[7]はオルトメトリクスの提唱者である Jason Priem らによって 2012 年 9 月より開始されたオルトメトリクス計測サービスである。Altmetric.com は論文単位の情報提供を行うことに焦点を当てているのに対し，Impactstory は著者単位の情報提供を行うことに焦点を当てている。2014 年 9 月より有償サービスを開始し，研究者の履歴書（Curriculum Vitae）の新たなスタンダードになることを目指し，活動している。

オープンアクセス出版の大手である PLOS（Public Library of Science）は，オルトメトリクスが提唱される前の 2009 年より，被引用数以外の指標を組み込んだ学術文献の影響度を計測するプロジェクト（PLOS Article-Level Metrics）に取り組んでいる [11]。ここでは，書誌情報（HTML）の表示回数や本文（PDF）のダウンロード回数など，出版者のみが参照できるデータも活用されている。

既存のオルトメトリクス計測サービスは，主に DOI を対象としてデータを収集しており，日本語の文献の大半が計測対象から漏れている。また，出版者や学術機関，開発者に対しては文献のランキング（注目されている文献一覧）が提供されているものの，一般の利用者には容易にアクセスできる状況ではない。つまり，ある文献の影響度を知ることはできても，影響度から文献を探すことはできず，学術文献になじみのない利用者に対し，学術文献に興味を持たせることは困難であった。さらに，各文献の影響度が総計としてのみ提示されている場合が多く，その経緯や詳細を知るのが難しいという問題も抱えている。筆者は，それらの問題を解決すべく，国産オルトメトリクス計測サービスの開発を行っている。

3. Ceek.jp Altmetrics

3.1　Ceek.jp Altmetrics の概要

Ceek.jp Altmetrics は筆者が開発および運営する国内唯一のオルトメトリクス計測サービスである。主に日本の学術文献のオルトメトリクスを計測することを目的として，2013 年 10 月 29 日にサービスを開始し [12]，2014 年 10 月現在，約 13 万件の学術文献に対し，約 41 万件の言及情報を蓄積している。筆者はソーシャルメディアで流通する学術文献に関するデータを収集し，新たな学術文献評価指標の研究開発を行っており，その成果として本サービスを提供している。

Ceek.jp Altmetrics は主に 2 つの機能を提供している。1 つ目の機能は，図 2 のようなランキング機能である。ここでは，急に言及されるようになった旬な学術文献情報を提供している。さらに，図 3 のように日別に集計した結果をカレンダー形式で表示する機能も用意している。これらの機能により，学術文献になじみのない利用者にとっても，学術文献を身近に感じて貰えると考えている。2 つ目の機能は，各文献の言及情報詳細である。ここでは，図 4 のように言及数のトレンドをみることができ，いつ流行したのか，あるいは，どれほどの期間流行したのかを容易に知ることができる。

図 2　ランキング機能

図 3　カレンダー機能

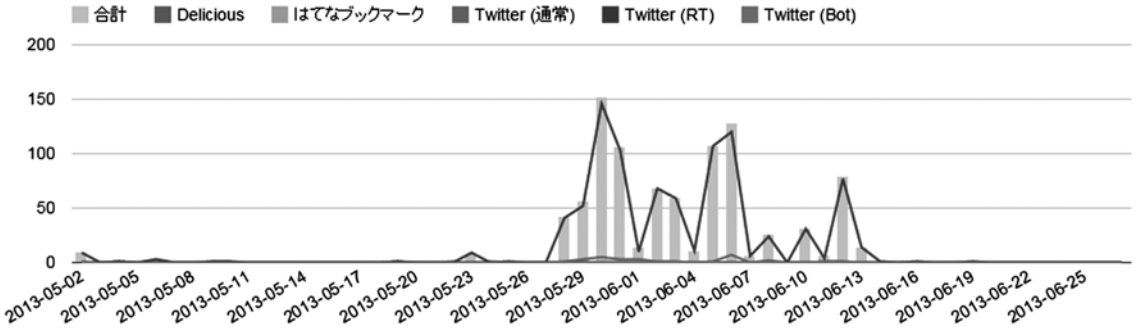


図 4　言及のトレンド情報

3.2　Ceek.jp Altmetrics の構成

図 5 に Ceek.jp Altmetrics のシステム構成を示す。本システムは，主にクローラ，データベース，データマイニング，ユーザインタフェースの 4 機能にわけることができる。クローラはソーシャルメディア等から言及情報を，学術情報サービス等から文献情報を収集し，データベースに格納する。データマイニングはデータベースに格納されたデータをもとにオルトメトリクスを算出し，データベースに格納する。ユーザインタフェースはアクセスもとのクライアントに応じた形式でデータベースに格納された情報を出力する。

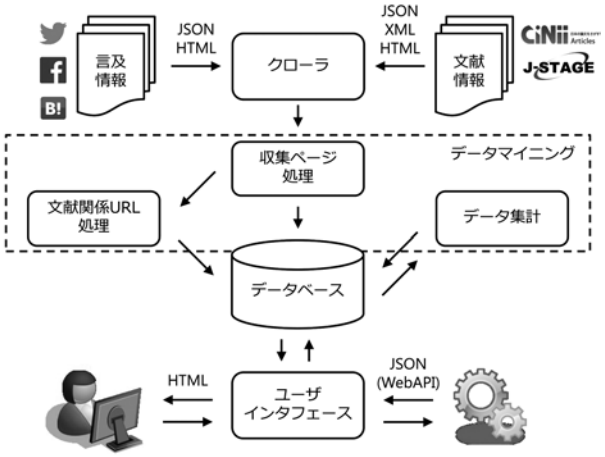


図 5　Ceek.jp Altmetrics のシステム構成

3.2.1　クローラ

クローラは言及情報収集クローラと文献情報収集クローラに大別でき，相互が繰り返して作動している。それぞれのクローラが収集対象とするウェブサイトはあらかじめ定められており，執筆時点においては，表 1 の通りである。言及情報収集クローラは主に学術文献の URL とそれに対する言及テキストを収集している。文献情報収集クローラは，言及情報収集クローラが収集した文献 URL と関連付けられる学術文献のメタデータを収集している。

収集対象となるウェブサイトが限定されているといえども，言及情報収集クローラがサイト内の全てのデータを収集することは困難である。本システムでは，収集対象ウェブサイトに存在する検索機能を利用し，効率的にデータを収集している。つまり，収集対象となる学術文献の URL の一部を検索し，その検索結果をもとに学術文献の URL とそれに対する言及テキストを収集する。例えば，CiNii に収録されている学術文献に対する，Twitter における言及を収集するケースを考える。CiNii に収録されている学術文献は「http://ci.nii.ac.jp/naid/110008898261」のような URL で提供されており，その URL には「ci.nii.ac.jp」が常に含まれている。Twitter から言及情報を収集する際は，公式に提供されている検索 API[13]に対し，「ci.nii.ac.jp」という検索クエリを送信し，その結果を収集している。

文献情報収集クローラは，後述するデータベースを参照しながら，未収集の文献情報を収集している。CiNii であれば，「http://ci.nii.ac.jp/naid/110008898261.rdf」などにアクセスすることにより，機械可読な形式で学術文献のメタデータにアクセスすることができ，文献情報に対するアクセスが容易である。

データを収集する際，可能な限り機械可読なページを収集しているが，JAIRO のように機械可読なページを提供していないケースもある。また，J-STAGE のように機械可読なページには十分なデータが含まれていないケースもある。それらのような場合，ウェブページをスクレイピング（解析）することで必要なデータを収集している。

表 1　収集対象ウェブサイト

言及情報収集クローラ		文献情報収集クローラ	
Facebook	ソーシャルメディア	AgriKnowledge	
Google+		CiNii	
Twitter		J-STAGE	
OKWave	FAQ	JAIRO	
Yahoo!知恵袋		情報学広場	
CiteULike	ブックマーク	国立国会図書館デジタルコレクション	
Delicious		JAIRO Cloud	機関リポジトリ
はてなブックマーク		一部の大学	
Wikipedia			
レファレンス協同データベース			

3.2.2　データベース

データベースには言及情報収集クローラが収集したデータ，文献情報収集クローラが収集したデータ，後述するデータマイニングによって処理されたデータが格納されている。データを格納するミドルウェアとして，MySQL[14]と Mroonga[15]を利用している。MySQL は著名なリレーショ

ナルデータベースであり，多くのウェブサービスで利用されている。しかし，現時点においては日本語を対象とした全文検索機能が不十分であるため，MySQL の全文検索機能を拡張する Mroonga も利用している。

本システムを構成するデータベースの ER 図（Entity Relationship Diagram）の抜粋を図 6 に示す。この図では，言及情報データとして Twitter のみを取り上げており，また，実システムにはそのほかのデータも含まれる。後述するデータマイニングで利用されるデータに絞り，その概要を述べる。まず，言及情報（Twitter）テーブルにツイート ID，言及者，言及テキスト，言及日時が格納される。そして，言及テキストから文献関係 URL を抽出し，URL テーブルにその URL を格納した上で，言及情報と URL との関係を保持する。URL テーブルでは，URL から文献を一意に特定する文献識別子を生成し，その情報をもとに文献情報データを収集した上で，文献情報テーブルに文献データを格納する。文献情報テーブルは国立情報学研究所が機関リポジトリ用に制定した junii2 フォーマット [16]に従ってスキーマを定義している。

3.2.3 データマイニング

データマイニング処理は，クローラの動作およびデータベースの管理と密接に関係している。ここでは，主に 4 種類のプログラムが稼働している。

(1)収集ページの処理

言及情報収集クローラおよび文献情報収集クローラが収集したページをデータベースに格納できる形式に変換する処理を行う。機械可読なページは XML や JSON で提供されており，それらのページを適切なライブラリを利用してパースする。HTML のような機械可読が難しいページは，あらかじめ定めたルールに従ってスクレイピング処理を行う。

(2)文献関係 URL の処理

言及テキストから学術文献に関係のある URL を抽出する。ここではテキストに含まれる URL のうち，収集対象文献サイトのホスト名を含む URL を抽出している。この際，URL の正規化は行わず，URL から文献を一意に特定する文献識別子を生成し，データベースに格納している。

文献関係 URL を正規化しない利点は，言及者による言及方法の詳細を調査できることにある。例えば，CiNii に収録されている「ソーシャルメディアの政治的活用 ：活用事例と分析事例から」に対する言及の明示は，次のようなものがあり得る。

- http://ci.nii.ac.jp/naid/110008898261
- http://ci.nii.ac.jp/naid/110008898261/ja/
- http://ci.nii.ac.jp/naid/110008898261/en/
- http://ci.nii.ac.jp/naid/110008898261.rdf

これらは，それぞれ異なる意図によって言及されたと推察でき，言及要因の分析を行う際に有用な情報となる。CiNii Articles の場合，「naid/」に続く 12 桁の数字によって文献を一意に特定できると考えられるため，先の URL 群から文献識別子を生成すると「naid:110008898261」となる。

(3)文献別言及数の集計

図 6 の通り，言及情報から文献情報まで関連付けることができるため，文献別の言及数を集計することができる。ここでは言及情報として Twitter のみを例示しているため，文献別集計データテーブルにも言及数（ツイート）しか明示していないが，実システムにおいては，Facebook やはてなブックマークなどのそのほかのサイトにおける言及数，さらにそれらの総計言及数，あるいは直近 1 日間の総計言及数などが格納されている。これらのデータは，図 2 のようなランキング機能に用いられる。

言及数を集計する際には，重複した言及を除去する必要がある。本システムでは，「ある言及者はある文献に 1 度のみ言及することができる」という制約を設け，集計の際には特定の言及者が多数言及したとしても，1 言及として集計している。これにより，ボット等による自動投稿の影響を軽減することができる。

(4)日別言及数の集計

本システムには，図 3 のように日別に集計した結果をカレンダー形式で表示する機能もある。この機能を実現するために，日別集計データテーブルを用意している。日別集計データテーブルには，ある日に最も言及された文献情報（文献識別子）とその言及数が格納されている。

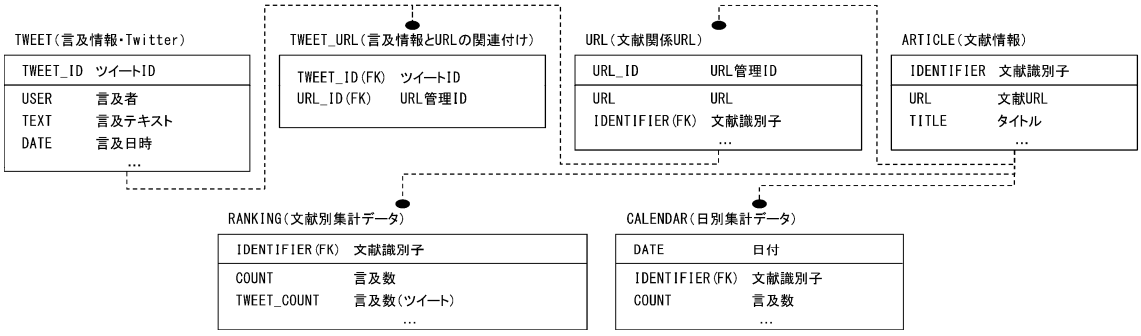


図 6 データベースの ER 図（抜粋）

3.2.4　ユーザインタフェース

ユーザインタフェースは，データベースを参照し，その時点における最新データを利用者に提供する。本システムでは，一般の利用者に最適化されたインタフェース（HTML）と機械可読なインタフェース（JSON）を用意している。

機械可読なインタフェースは Ceek.jp Altmetrics API[17] を通じて提供される。このようなインタフェースは Web API[18]と呼ばれ，外部システムとの連携を容易にする。本システムが提供する Web API は，先行する Altmetric.com が提供する Web API の機能 [19]と互換がある。互換を持たせることで，外部システムの開発者は，Altmetric.com のデータにアクセスするのと同様に，Ceek.jp Altmetrics のデータにアクセスすることができる。

3.3　データの収集状況

本システムは，2013 年 4 月より開発を始め，2013 年 10 月 29 日にリリースした。開発開始以前の言及情報も収集し（Facebook は技術的な制約があり，開発開始以降の言及情報のみ収集），2014 年 10 月 31 日現在，約 13 万件の学術文献に対し，約 41 万件の言及情報を蓄積している。学術文献の内訳は，CiNii Articles が 65%，国立国会図書館デジタルコレクションが 13%, J-STAGE が 8%と続く。また，言及情報の内訳は，Twitter が 84%，はてなブックマークが 9%, Wikipedia が 5%と続く。Twitter の収集データをさらに細かくみると，37%が自動投稿（ボット）である。これは CiNii ウェブ API コンテストで作成された「論文ったー」[20,21]の影響が大きい。なお，各文献に対する言及数を調べたところ，図 7 のようにジップ分布に従うことが確認された。同様に，各言及者の言及数もジップ分布に従う。

先に述べたように，収集した学術文献の 65%が CiNii Articles に収録されている文献である。しかし，この比率の傾向は，言及サイトによって異なる。図 8 は言及情報を比較的多く収集できた 3 サイト，Twitter，はてなブックマーク，Wikipedia における，各学術文献の出現分布である。Twitter やはてなブックマークにおいては CiNii Articles の文献が多数言及されているものの，Wikipedia においては国立国会図書館デジタルコレクションの文献が多数言及されている。Wikipedia においては，古典資料が好まれる傾向があると考えられる。

今後もオルトメトリクスが有効に機能するためには，言及データが増加している必要がある。仮に減少傾向であるならば，近い将来，言及情報を得られなくなることから，その情報を文献評価指標として利用するのは困難になる。図 9 は主要な言及サイトにおける，言及数の伸びである。縦軸は月間の言及数を表している。言及数の多い Twitter においては継続的に上昇傾向であるものの，そのほかのサイトにおいては横ばいあるいは減少傾向がみられ，今後も注視していく必要がある。

4．今後の展望

現状，Ceek.jp Altmetrics も含め，オルトメトリクス計測サービスは言及数の重み付け集計数をオルトメトリクスとしており，媒体が持つ情報を十分に生かしていない。例えば，Twitter のデータには，言及テキスト以外にも，言及者同士の関係（ソーシャルグラフ）が含まれる。このような情報を有効に活用することで，より実態にあった指標を開発できる可能性がある。図 10 は，ソーシャルグラフのデータを用い，言及の伝搬経路を推定した 2 つのネットワークである（太い矢印は最初の言及者）。いずれも言及数が同じであるものの，(a)は遠くのユーザまで伝搬している一方，(b)は中心ユーザから近いユーザにまでしか伝搬していない。これらの違いを指標として表現する必要性を感じ

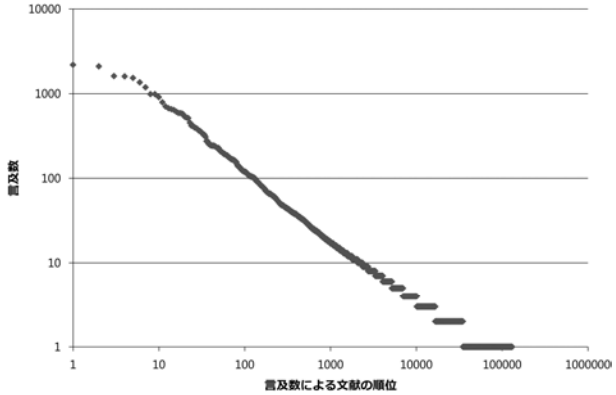


図 7　文献に対する言及数とその順位の関係（ジップ分布）

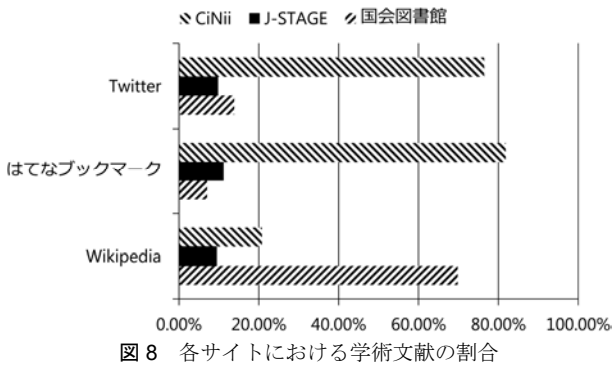


図 8　各サイトにおける学術文献の割合

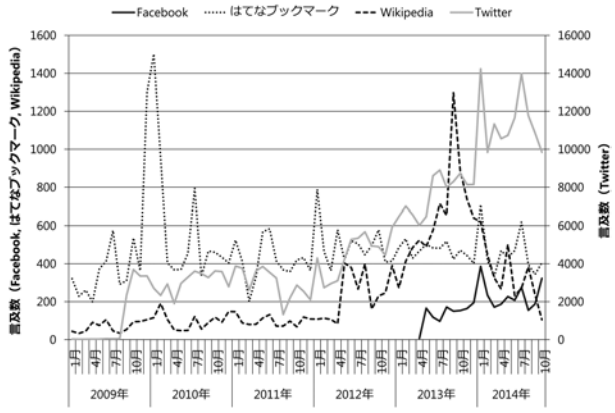
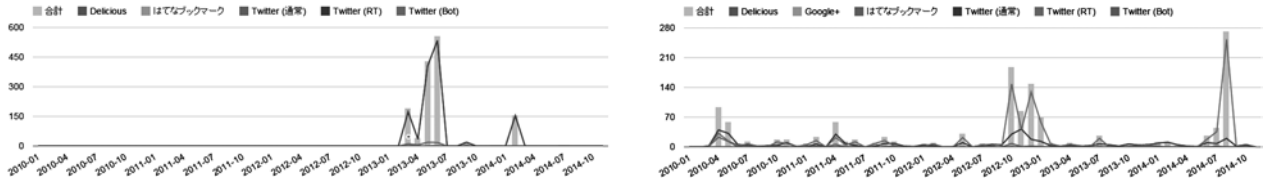
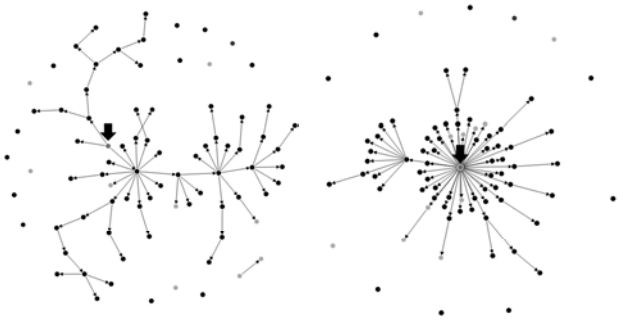


図 9　主要なサイトにおける言及数の月別変動



(a) (b)

図 11　月間言及数の変動の異なり



(a) (b)

図 10　情報伝搬経路の異なりに対する印象の違い

ている。

収集している言及情報には,言及日時が付与されている。時間の情報を考慮することによっても，より良い指標を開発できる可能性がある。図 11 はほぼ同数の言及がある 2 文献のそれぞれの月間変動グラフである。(a)は短期間で言及が収束しているが，(b)は長期間にわたって言及が継続している。これらの違いを指標として表現できると有用であると考えている。

オルトメトリクスの提供以外に目を向ければ,そもそも，ソーシャルメディアで言及される学術文献はどのような文献であるか，という調査研究も必要であると考える。この調査研究は，社会的な影響の大きい学術文献の特性を明らかにしようとするものであるが，従来の調査では，大半は PubMed を中心とする医療に関する文献の分析に偏り [9], 言語も英語に限定される [22,23]。例えば医療に関する文献は自身の生活（健康）に密接に関わる一方，歴史学に関する文献は医療よりも生活に遠い存在であると考えられる。オルトメトリクスはインパクトファクターと異なり，分野を横断しての評価が可能であるとされているものの [3]，分野を横断した定量的な調査が行われておらず，適切に検証されていない。本システムによって蓄積したデータを利用し，分析を進めたいと考えている。

Ceek.jp Altmetrics で収集したデータは，各ウェブサービスから自動収集したものである。収集システムを開発するコストが存在するものの，一度開発しさえすれば，全自動で収集および分析が行える。つまり，運営自体に大きなコストがかかっていない。今後も安定的な運営を続けるとともに，Web API の提供はもちろんのこと，収集したデータをオープンにし，計量書誌学等に関係する研究者が利用できるような基盤システムにしていきたいと考えている。

注・参考文献

（Web 参照日は全て，2014 年 10 月 31 日）

1) Eugene Garfield. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. Science, 1955, vol.122, no.3159, p.108-111.
2) J. E. Hirsch. An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 2005, vol.102, no.46, p.16569-16572.
3) 林和弘．科学技術動向研究　研究論文の影響度を測定する新しい動き：論文単位で即時かつ多面的な測定を可能とする Altmetrics．科学技術動向，2013，no.134，p.20-29.
4) Paul Mcfedries. Measuring the Impact of Altmetrics. IEEE Spectrum Magazine, 2012, vol.49, no.8, p.28.
5) Finbar Galligan, Sharon Dyas-Correia. Altmetrics: Rethinking the Way We Measure. Serials Review, 2013, vol.39, no.1, p.56-61.
6) Altmetric http://www.altmetric.com/
7) Impactstory https://impactstory.org/
8) Ceek.jp Altmetrics http://altmetrics.ceek.jp/
9) Jennifer Lin, Martin Fenner. Altmetrics in Evolution: Defining & Redefining the Ontology of Article-Level Metrics. Information Standards Quarterly, 2013, vol.25, no.2, p.20-26.
10) David Shotton. CiTO, the Citation Typing Ontology. Journal of Biomedical Semantics, 2010, vol.1 no.S1.
11) Jennifer Lin, Martin Fenner. The Many Faces of Article-Level Metrics. Bulletin of the American Society for Information Science and Technology, 2013, vol.39, no.4, p.27-30.
12) 日本の論文の Altmetrics 計測サービス，Ceek.jp Altmetrics が公開 http://current.ndl.go.jp/node/24696
13) GET search/tweets | Twitter Developers https://dev.twitter.com/rest/reference/get/search/tweets
14) MySQL http://www-jp.mysql.com/
15) Mroonga http://mroonga.org/ja/
16) メタデータ・フォーマット junii2 http://www.nii.ac.jp/irp/archive/system/junii2.html
17) Ceek.jp Altmetrics API http://api.altmetrics.ceek.jp/
18) 高久雅生．Web API の過去・現在・未来．情報の科学と技術，2014，vol.64，no.5，p.162-169.
19) Altmetric API documentation http://api.altmetric.com/
20) 論文ったー https://twitter.com/ronbuntter
21) 山田俊幸．空気を読んで論文を紹介する「論文ったー」．専門図書館，2012，no.255，p.27-33.
22) Ludo Waltman, Rodrigo Costas. F1000 Recommendations as a Potential New Data Source for Research Evaluation: A Comparison With Citations. Journal of the Association for Information Science and Technology, 2014, vol.65, no.3, p.433-445.
23) Mike Thelwall, Stefanie Haustein, Vincent Larivière, Cassidy R Sugimoto. Do Altmetrics Work? Twitter and Ten Other Social Web Services. PLOS ONE, 2013, vol.8, no.5.

Special feature: Beyond Bibliometrics. New Challenge for Bibliometrics －Development of the Altmetrics Measurement Service－. Mitsuo Yoshida (Toyohashi University of Technology, 1-1 Higarigaoka, Tempaku-cho, Toyohashi, Aichi, 441-8580 JAPAN)

Abstract: Academic information has been evaluated by citation-based indicators such as the Impact Factor and *h*-index. Since around 2010, mentions in social media have been used instead of citation-based indicators. These indicators are called “Altmetrics”. This paper discusses the overview of altmetrics and the national altmetrics measurement service “Ceek.jp Altmetrics” that the author has been developing. In addition, the paper reports the trend of mentions for Japanese academic information in social media and discuss the future outlook.

Keywords: Academic information / Altmetrics / Social Media / Crawler / Web API