

# オープンソースによるイノベーション普及過程 探索ツールの作成\*

河 合 勝 彦<sup>†</sup>  
櫻 井 雄 大<sup>‡</sup>

## 1. はじめに

本稿は、社会におけるイノベーションや新製品の普及過程をモデル化するために必要な、Web上のデータを収集するツールを紹介する。なお、物理的に形をもたないモノの普及を考える場合には、「新製品」という言葉を「流行」、「ニュース」等に置き換えれば良い。普及過程のモデル化とは、あるイノベーションもしくは新製品（流行）が、ある期間内に、どのような過程を経て人々の間に普及していくかを、仮説として示すものである。

イノベーション、新製品、流行等が、どのように社会の人々（消費者）に普及していくかを解明することは、企業の経営戦略や政府の産業政策等にとって非常に重要な問題である。実際、社会科学（マーケティング、経済学等）の分野では、こうした普及現象をモデル化する試みが従来から地道に続けられている。その代表的なものとして、社会における普及の拡散状況を微分方程式でモデル化する、イノベーション普及モデル（以下、普及モデル）が存在する [10]。

数学的に洗練されてはいるが、データの入手が困難であるという理由から、普及モデルを実際にデータで検証することは案外難しい。例えば、耐久消費財を分析の対象とする従来の普及モデルは、国などが提供する少品目の耐久消費財普及の調査データを、その検証に利用するのが常である。その一方、産業社会の多くの部分を占めるサービスの普及、およびアイデアや概念の普及（流行）をモデル化するためには、各種業界団体による未・非公開データ等に頼らなければいけない場合が多い。しかしながら、爆発的に増加しつつあるWeb上の情報に着目すれば、こうした実体を持たないモノも含めた普及現象について観察することが、ある程度容易になる。

ただし、Web上の情報は非定型であることが多いので、それを調査・分析に利用するために

---

\* 本稿は、科学研究費補助金（課題番号 19530388）の助成を受けて行った研究成果の一部である。

† 名古屋市立大学大学院経済学研究科 E-mail: kkawai@econ.nagoya-cu.ac.jp

‡ 名古屋市立大学大学院経済学研究科博士後期課程 E-mail: sakurai.yuta@gmail.com

は、その情報を利用可能なものに加工する必要がある。具体的には、Web 上から目視以外の効率的な方法で、利用したい特定の部分の情報を取り出す方法が必要である。さらに、その情報を使って統計分析をしたり、グラフ等で可視化したりすることが可能なように、それを定型的なインプットデータに加工する必要がある。

こうした問題点に鑑み、われわれは、Web 上の必要なデータを収集し、インプットデータとして再加工するツールとして、2つのプロトタイププログラム、SWAT\_CI と XpathChecker を作成した。本稿は、この2つのツールを作成した過程について詳述する。ただし、本稿は、これらのツールの基本性能および活用法の紹介を目的としたものであり、われわれの実際の調査・研究にどのように利用されたかということについては、また稿を改めて紹介したい。

論文の構成は次の通りである。まず第2章では、SWAT\_CI について解説する。次に、第3章では、XpathChecker について説明する。第4章では、両ツールの活用方法についていくつかの簡単な実例をあげる。そして、第5章では、今後の課題を提示し、まとめをおこなう。

## 2. SWAT\_CI について

本章では、SWAT\_CI (Social Web Analysis Tools built by CodeIgniter) の仕様について解説する。SWAT\_CI は、Web 上における普及メカニズムの調査を目的として開発されたオープンソースソフトウェア<sup>1)</sup> である。

SWAT\_CI は、ネットワーク上の blog が提供するサービスの更新情報を定期的に巡回し、blog 執筆者ごとに、そのソーシャルブックマーク<sup>2)</sup> 数、RSS フィード<sup>3)</sup> 購読者数、更新回数、blog 記事内容等を収集する。さらに、収集されたデータや形態素解析<sup>4)</sup> されたテキスト内容の統計分析を行い、様々な方法で図示およびデータ出力する。

### 2.1. ツール作成の目的

本プログラムは、blog を介した、ネットワーク上の人のつながりを分析<sup>5)</sup> することを目的に

---

1) そのソースコードが公開されており、誰でも自由に再利用および再配布することができるソフトウェア。厳密な定義については、『オープンソースの定義 [1]』を参照のこと。

2) ソーシャルブックマークとは、ブックマークをネットワーク上で不特定多数の人々と共有可能にする仕組み。付加的な機能により、自分のブックマークが、どの程度、他人と類似しているかということや、任意の Web ページがどのようにブックマークされているかという傾向(認知度、人気度)を知ることが可能になる場合もある。

3) ある Web サイトにおける文章の見出しや内容の要約を XML 形式で提供するもの。

4) 文章を構成する最小の意味単位を形態素と呼ぶ。ある文章に対して、この形態素に分解する作業をおこなうことを、形態素解析と呼ぶ。

作成されている。具体的には、以下の機能要件を満たすべく開発が進んでいる。

1. blog の時系列データ（テキスト、コメント、被ブックマーク等）の収集。
2. 収集されたデータおよび分析結果の「見える化」（グラフ等による出力）。
3. 過大な専門知識を必要としない、目的に応じた柔軟な分析設定。
4. 様々なネットワーク分析手法のライブラリ化<sup>5)</sup>。

## 2.2. 動作環境

本プログラムは、定期的なデータ収集が最重要目的の一つであることから、常時稼働を前提とするサーバ環境が必要である。さらに、実行するサーバにおいては、以下の要件を満たす必要がある。

- ・ インターネットに接続されており、Web ページの取得が可能である。
- ・ PHP5<sup>7)</sup> 以上がインストールされ、PHP スクリプトが実行可能である<sup>8)</sup>。
- ・ PHP が動的に実行可能な HTTP デーモン<sup>9)</sup> が導入されている。
- ・ RDBMS<sup>10)</sup> が動作し、データベースの作成と利用が可能である。
- ・ 統計分析ツール R<sup>11)</sup> がインストールされ、R スクリプトが実行可能である。

なお参考として、以下に開発時の実行環境を提示する。

- ・ OS : Ubuntu Linux 8.04 LTS<sup>12)</sup>
- ・ PHP 5.2.5
- ・ Apache 2.2.8 (Unix)
- ・ MySQL 5.0.51a

---

5) こうした Web 上における人と人のつながりを分析する方法として、物理学およびコンピュータ科学の分野では、複雑ネットワーク分析というものが用いられている。われわれのツールも同分析の手法を多数取り入れている。この分析手法の詳細については、今野・井手 [4] および増田・今野 [6] が参考になる。

6) 高度なネットワーク分析を可能にするツールとして、既に NetworkX [12] のような優れたものが存在する。その一方、本ツールは、データ収集の機能をメインとして設計されている。

7) PHP は、Web ページ埋め込み型のスクリプト言語である。PHP 言語については、その公式サイト [13] が参考になる。

8) Web アプリケーションの効率的な製作には、フレームワークの利用が必須である。今回のアプリケーション作成には、CodeIgniter [2] を利用した。

9) Apache [9] もしくは同等の機能を持つもの。

10) Relational Database Management System (リレーショナルデータベース管理システム) のこと。MySQL [11] もしくは同等の機能を持つものが必要。

11) R の詳細については、公式サイト [14] を参照のこと。

12) Linux ディストリビューションのひとつ。非常に操作性に優れ、サポート体制が良いことに定評がある。詳細については、公式サイト [16] を参照されたい。

## 2.3. 機能

本プログラムの提供する機能の概要は、以下の通りである。

- ・収集対象 blog サービスの管理：データ収集対象となる blog サービスを登録・修正・削除する機能。
- ・データ収集機能：上の収集対象 blog サービス管理機能により登録された収集対象について、定期的に blog 更新情報を執筆者別に収集し、データベースに保存する機能。
- ・データ分析機能：収集されたデータを対象として、ユーザの操作に従って、統計分析・形態素解析を行い、その結果を表示あるいはデータベースに保存する機能。
- ・データ出力機能：収集・分析されたデータを、ユーザの操作に従ってブラウザ上に表示するか、もしくは種々のファイル形式として出力する機能。

なお、上に挙げた機能のいくつかは、まだ実装されておらず<sup>13)</sup>、オープンな体制での開発が継続しておこなわれている。

## 2.4. 処理の概要

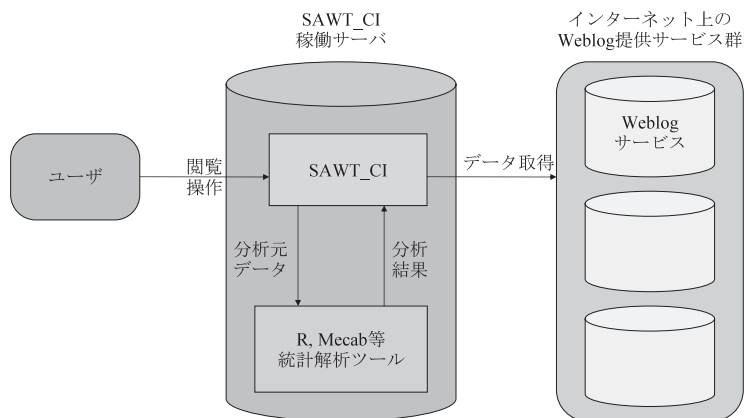


図1 SWAT\_CIの動作概念図

(出所：筆者作成)

13) 未実装の項目について、以下に列挙する。

収集対象 blog サービス管理：対応している blog サービスは、「はてな [5]」の「はてなダイアリー (<http://d.hatena.ne.jp/>)」のみとなっており、他サービスには対応していない。

データ収集機能：定期収集機能に問題があり、特定状況下ではデータの収集が正常に行われない。

データ分析機能：取得したデータの形態素解析部分が不完全である。

データ出力機能：グラフ等での図示、各アプリケーションに対応したファイル形式での出力機能が実装されていない。

データ公開機能：公開ポリシーの設定・適用が実装されていない。

本プログラムは、Web サービスとして動作する。利用ユーザは、自身のアカウントでログインし、収集対象の Blog サービスについて次の項目を登録する。

1. blog サービス名
2. blog サービス URL
3. 備考

そして、登録された内容はサーバ内のデータベースに保存される。

次に、本プログラムはプログラム設置時に設定された周期ごとに、以下の動作を行う。

1. 登録された blog サービスをチェックし、その更新情報 (RSS フィード) を取得する。
2. 更新情報から得られる執筆者・記事内容をデータベースに保存する。
3. blog サービスによって提供される追加情報 (ソーシャルブックマーク数等) を取得する。

ユーザはこれらの取得済みデータについて、ブラウザ上の操作によって、統計量の計算やデータを視覚化したグラフ等の出力を行うことができる。例えば、統計処理関係の操作がリクエストされた場合、本プログラムはその処理に対応する R スクリプトを選択し、収集済データと合わせて R に送る。そして、結果として得られる R からの出力結果 (解析結果の文字列、グラフ画像等) をブラウザ上に出力する。

### 3. XpathChekcer について

本章では、XpathChecker の仕様について解説する。前章で説明した SWAT\_CI と同様に、XpathChecker は、Web 上における普及メカニズムの調査データ収集を目的として開発されたオープンソースソフトウェアである。

本プログラムは、普及現象の解明にとって重要な、ネットワーク上における人のつながりを分析することを目的として作成されている。重要な機能としては、特定の URL で指定された Web ページを定期的に巡回し、XPath<sup>14)</sup> で指定された要素のテキストデータを収集する<sup>15)</sup>。さらに、収集されたデータや形態素解析されたテキスト内容の統計分析を行い、様々な方法で図示およびデータ出力をおこなう。

---

14) 本稿では、XPath の仕組みを詳述することはしない。XPath の仕様については、その標準化団体 W3C の Web サイト [17] を参照のこと。また第 4 章に簡単な利用例を掲載する。

15) 類似の Web サービスとして、Kayac (<http://www.kayac.com/>) の手による、XPathGraph [8] が既に存在する。ただし、XPathChecker は、研究用にオープンソースとして一般に公開されているという点に特徴がある。さらに、XpathChecker は、社会科学分野で学術的な利用が可能なように統計分析の機能を取り入れる予定であるという点で XpathGraph とは相違している。

### 3.1. ツール作成の目的

本プログラムは、Web 上のデータをもとにして、人と人のつながり（ソーシャルネットワーク）を分析のすることを目的としたツールである。この人と人のつながりを分析することにより、普及現象を解明する手助けとなる。

具体的には、以下の要件を満たすことを目的とする。

1. 特定のフォーマットに左右されない、任意の Web ページの時系列データ収集。
2. XPath クエリによる柔軟なデータ収集設定。
3. 収集されたデータおよび分析結果の「見える化」（グラフ等による出力）。
4. 様々な分析手法のライブラリ化。

### 3.2. 動作環境

定期的なデータ収集が、本プログラムの重要な目的の一つであることから、常時稼働を前提とするサーバ環境が必要である。

なお、本プログラムを稼働させるサーバの要件および開発時の環境は、SWAT\_CI と同じである。

### 3.3. 機能

本プログラムの提供する機能の概要は、以下の通りである。

- ・収集対象 URL, XPath クエリ管理：データ収集対象となる Web ページ (URL)、およびそのデータのページ内における位置を一意に決定する XPath クエリを、登録・修正・削除する機能。
- ・利用ユーザ管理：複数人での共同利用を可能とするための利用ユーザの登録・修正・削除、および前述の収集対象 URL 等とユーザの紐づけを行う機能。
- ・データ収集機能：前述の収集対象 URL, XPath クエリ管理機能により登録された収集対象を、設定内容に従って定期的に収集し、データベースに保存する機能。
- ・データ分析機能：収集されたデータを対象として、ユーザの操作に従って統計分析・形態素解析を行い、その結果を表示あるいはデータベースに保存する機能。
- ・データ出力機能：収集・分析されたデータを、ユーザの操作に従ってブラウザ上に表示、あるいは種々のファイル形式として出力する機能。
- ・データ公開機能：収集・分析されたデータを、あらかじめ設定された公開ルールに従ってユーザ・外部に公開する機能。

なお、上に挙げた機能のいくつかは、まだ実装されていないが<sup>16)</sup>、オープンな体制での開発が継続しておこなわれている。

### 3.4. 処理の概要

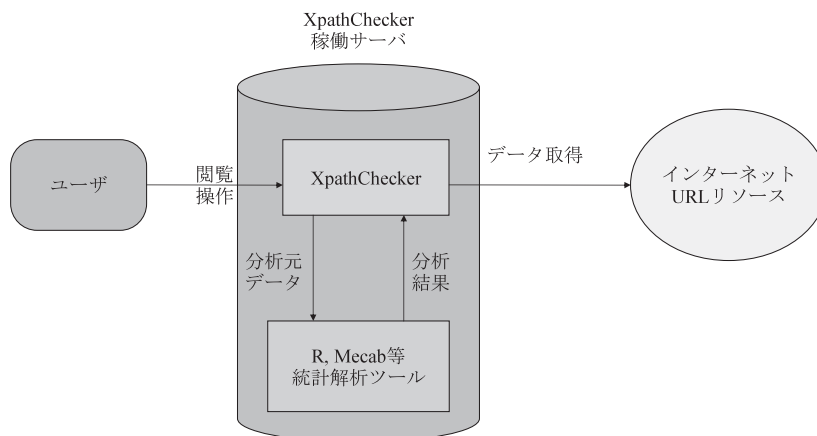


図2 XpatchChecker の動作概念図

(出所：筆者作成)

本プログラムは、Web サービスとして動作する。ユーザは自身のアカウントでログインし、収集の対象となるデータについて、以下の項目を登録する。

1. データ名
2. 対象 URL
3. 対象 XPath
4. データ種別（数値、文字列）
5. 公開範囲
6. 備考

なお、登録された内容はデータベースに保存され、必要時の再利用が可能である。

16) 機能未実装の項目について、以下に列挙する。

利用ユーザ管理：現状では、ユーザごとに収集対象の追加・編集・削除等の権限を割り当てる機能が実装されていない。

データ収集機能：XPath で指定された Web ページの部分参照が認識されない場合がある。定期収集機能に問題があり、特定状況下ではデータの収集が正常に行われない。

データ分析機能：取得したデータの形態素解析部分が不完全である。

データ出力機能：グラフ等での図示、各アプリケーションに対応したファイル形式での出力機能が実装されていない。

データ公開機能：公開ポリシーの設定・適用について未実装となっている。



さらに、本プログラムは、設置時に設定された一定の周期ごとに、次の動作を行う。

- ・登録された収集対象データをチェックし、対象 URL の Web ページを取得する。
- ・対象 XPath で一意に定まる HTML の要素を取り出し、その内容をデータベースに保存する。

SWAT\_CI と同様に、ユーザは、これらのデータについて、ブラウザ上から各種統計量を計算したり、結果をグラフィカルに表示することができる。例えば、統計処理関係の操作が行われた場合、本プログラムは処理に対応する R スクリプトを用い、先の収集済データと合わせて R に送る。その結果として得られる R からの出力結果（解析結果の文字列、グラフ画像等）をブラウザ上に出力する。

## 4. 活用例の考察

本章では、SWAT\_CI と XPathChecker の潜在的な活用例について考察を加える。

なお、2009 年 6 月 1 日現在、これら 2 つのツールはまだ実験運用中であり、完全に稼働しているとは言い難い。よって、実際の活用例というよりも、今後の活用可能性に重点を置いて説明を加えているということに留意して欲しい。

### 4.1. SWAT\_CI の活用例の検討

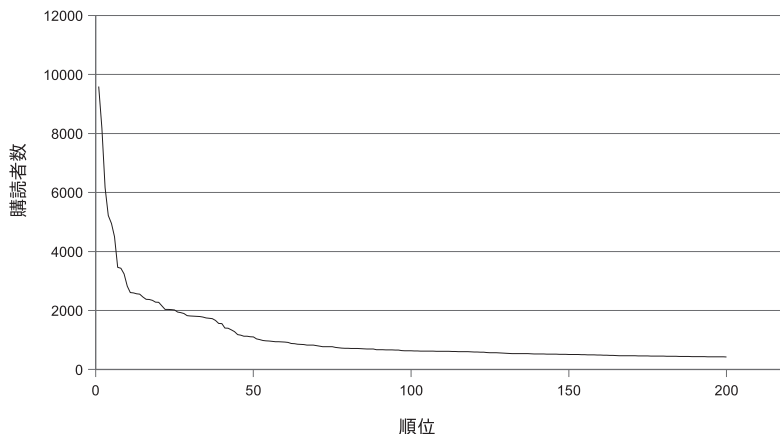


図3 blog の購読者数と順位  
(出所：TopHatenar [7] から入手したデータをもとに筆者作成)

まず、SWAT\_CI の活用方法について考える。例えば、SWAT\_CI を用いて Web 上の blog 記事を多数収集することにより、ある事柄の流行が、ある blog から他の blog へどのように広がるかを分析および視覚化することが可能になる。



さらに、どの blog が、どの程度に人気があるのかということをグラフ化する例を考えてみよう。SWAT\_CI は、そうした集計および表示機能を持っている。図 1 は、はてなダイアリー<sup>17)</sup>における購読者獲得数（被ブックマーク数）のうち、上位 200 人のデータをグラフ化したものである。y 軸には、あるブログの購読者数を、x 軸には、その購読者数の全体における順位をプロットしている。このグラフの形状からわかるように、blog の購読者数には、近似的にスケールフリー性<sup>18)</sup>が観察できる。なお、このようなネットワーク構造は、人間社会の多くの場面で確認できる。こうしたスケールフリー性のネットワーク構造における流行の形態は、ランダムなつながりのネットワークとは、かなり違ったものとなるだろう<sup>19)</sup>。

また簡単な別の例としては、形態素解析を用い、ある期間における特定単語の出現量を blog 執筆者ごとに調べることで、その単語がいつ頃から出現し始めたか、それが周りの記事にどれだけ広まったかなどを調査することができる。Web ページ上に、メタデータ（タグクラウド<sup>20)</sup>等）を自動生成するといった応用も容易であろう。

## 4.2. XpathChecker 活用例の検討

次に、XPathChecker の活用例を考える。SAWT\_CI は、データ収集の対象を、Web 上の Blog と呼ばれる領域に限定することにより、広範囲のネットワークを観測対象としていた。その一方、XPathChecker はある特定の 1 ページを定点観測することで、ネットワークの 1 ノードの発信する情報を詳細に調べることができる。なお、XPath を用いることにより、更新情報（RSS フィード）の発信を行っていない Web ページであっても、手軽にその更新を確認することができる。

例えば、(a)政府、自治体等の広報サイトと(b)その影響範囲内にあるであろう任意の Web ページを XPathChecker の収集対象とすることで、(a)が(b)に与える影響の程度や、その発生が確認できるまでの時間を測定できる。つまり、XPathChecker は、Web ネットワークの任意の 2 点間の関係や認知的な距離といったものを調べる用途に適している。

---

17) 脚注 13 でも説明しているが、「はてな」における blog サービス。

18) スケールフリー性とは、ネットワークのつながりの次数分布に特徴を見出したものである。数学的に説明すると、比例定数を  $\alpha$ 、べき指数を  $\gamma$  とした時、ノードの次数が  $k$  となる確率  $p(k)$  が  $p(k) = \alpha k^{-\gamma}$  となるような状態、つまり次数分布がべき分布であることを表す。スケールフリーな構造とは、現実の人々の交友関係を例に取ると、友人が多くいる人が少し存在し、他の大多数はそれと比べて非常に少ない友人しかいないという状況にあてはまる。

19) 河合 [3] は、この 2 つのネットワークにおける普及の形態の違いを簡単なシミュレーションモデルで示している。

20) タグと呼ばれる区分情報を、視認性が良くなるように、その人気に比例させて形状を大小に変化させたもの。

表1 Twitter における following と follower の関係

<i>user_id</i>	<i>following</i>	<i>follower</i>
4	1035	1078
14	1400	896
11	2001	874
18	1074	725
34	903	627
1	1268	368
5	626	366
15	289	313
24	263	274
3	335	256

(出所: Xpath Chcker により筆者が作成)

また、表1は、Web 上にある数値データ収集の一例として、ある Twitter<sup>21)</sup> ユーザがフォローしているユーザに対して、following (ユーザがフォローしている数) および follower (ユーザがフォローされている数) のデータを取得したものである。XPathChecker におけるデータ取得の構文は次の通りである：

- ・ URL : `http://twitter.com/ [ユーザ名 (user_id)]`
- ・ XPath : `//*[@id = "following_count"]`

このようなデータを時系列で取得することによって、任意のユーザの、ネット上における(発言の)影響力等を分析することが可能になる。

## 5. 今後の課題とまとめ

本稿は、普及現象について調査および研究をするために役立つ、Web 上のデータを収集し分析するツール、SAWT\_CI および XpatchChecker の2種類を紹介した。

残念ながら、解説を加えたすべての機能が現時点で実現されているわけではない。もちろん、研究ツールとして最低限の利用は可能な状態ではあるが、プログラミングや分析方法の専門知識を持たない人々の利用は難しい状況である。従って、今後も地道に開発を続けていくことが必要である。

さて、われわれは、SAWT\_CI および XpatchChecker の開発にあたり、オープンソースとい

21) 最近、インターネット上で爆発的に流行しているマイクロ blog と呼ばれる Web サービス。ユーザは、他の任意のユーザをフォローする (following) ことによって、そのユーザの「つぶやき (投稿)」を読むことが可能にある。また、フォローされる (follower) ことによって、自分の「つぶやき」を、そのフォロワーに示すことが可能になる。

う開発方法を採用した。オープンソースによる開発によって期待されることは、プログラムの不具合の発見や追加機能の要望等のフィードバックが迅速に得られるということである。また、オープンソースのライセンスにより、この2つのプログラムを誰もが自由に改変することができるし、再配布も可能である。内外からの、忌憚のないアドバイスおよび意見を期待したい。

今後の課題としては、まず、すべての仕様機能を実現化し、不具合（バグ）をなるべく少なくすることを挙げておく。また、工学的なテクニックだけではなく、社会科学分野におけるアカデミックな議論にもとづいた、重要なネットワークアルゴリズムや分析手法の機能追加をおこなっていきたい。

## 補論 ソースコードの入手について

本稿で紹介した SAWT\_CI および XpathChecker は、いずれも自由に再利用が可能なフリー・オープンソースソフトウェアとして公開されている。第3者が容易にアクセスできるように、ソースコードの概略および掲載場所については以下の URL に記述している。

・ <http://diffusion.kklab.info/>

2009年6月1日現在、SAWT\_CI の総行数は 3,079、そして、XpathChecker の総行数は、4,159 である。なお、原則、第3者から提供されたライブラリの行数は除外しているが、コメントと空白行は含んでいる。

## 参考文献

- [1] オープンソースの定義  
<http://opensource.jp/osd/osd-japanese.html>  
(2009年6月1日採録)
- [2] 河合勝彦, 鈴木憲治, 安藤建一 (2008)  
『CodeIgniter 徹底入門』 翔泳社.
- [3] 河合勝彦 (2009), イノベーション普及モデル  
の再考—サービス普及のモデル化を中心にして—,  
*Discussion Papers in Economics, Society of Economics*, Nagoya City University, vol499,  
2009, 1-10.
- [4] 今野紀雄, 井手勇介 (2008) 『複雑ネットワーク  
入門』 講談社.
- [5] はてな, <http://www.hatena.ne.jp/> (2009年6  
月1日採録)
- [6] 増田直紀・今野紀雄 (2005) 『複雑ネットワーク  
の科学』 産業図書.
- [7] TopHatenar, <http://tophatenar.com/> (2009年  
6月1日採録)
- [8] XPathGraph, <http://xpath.kayac.com/> (2009  
年6月1日採録)
- [9] The Apache Software Foundation,  
<http://www.apache.org/> (2009年6月1日採録)
- [10] Bass, F. M. (1969) "A New Product Growth  
Model for Consumer Durables," *Management  
Science*, Vol. 15, pp. 215-227.
- [11] MySQL: The world's most popular open  
source database, <http://www.mysql.com/> (2009  
年6月1日採録)
- [12] NetworkX, [http://networkx.lanl.gov/index.  
html](http://networkx.lanl.gov/index.html) (2009年6月1日採録)

- [13] PHP: Hypertext Preprocessor, <http://www.php.net/> (2009 年 6 月 1 日採録)
- [14] The R Project for Statistical Computing, <http://www.r-project.org/> (2009 年 6 月 1 日採録)
- [15] Twitter: What are you doing ?, <http://twitter.com/> (2009 年 6 月 1 日採録)
- [16] Ubuntu Home Page/Ubuntu, <http://www.ubuntu.com/> (2009 年 6 月 1 日採録)
- [17] XML Path Language (XPath), <http://www.infotaria.com/jp/contents/xmldata/REC-xpath-19991116-jpn.htm> (2009 年 6 月 1 日採録)

(2009 年 6 月 16 日受領)