O-004

セマンティック Web における効率的なメタデータ収集法の評価

Performance Study Approaches to Collecting Metadata over Semantic Web

及川 啓† 児玉 英一郎† 王 家宏† 高田 豊雄†
Kei Oikawa Eiichiro Kodama Jiahong Wang Toyoo Takata

1. はじめに

近年のインターネットの普及によって，Web ページは増加の一途をたどり，現在の Web ページ総数は 80 億 URL 以上にもなっている．このため利用者は自分の目的に合った Web ページを直接探すことが困難になっている．そこで利用者は，目的の Web ページを発見するために，Yahoo や Google といった検索エンジンを利用するのが一般的である．この検索エンジンの利用により，利用者は，直接探すよりは容易に目的の Web ページを発見できる．しかし，現在の検索エンジンには，トピックドリフト問題[1]という良く知られた問題があり，これによって利用者は，検索する際に時間的及び精神的負担を強いられており，十分満足のいく状態にはなっていない．

このトピックドリフト問題は，一般的には検索エンジンの問題として捉えられ，表層的に見ると，現在の検索エンジンが検索語と Web ページ内のテキストとのテキストマッチングしか行っておらず，Web ページ内の語彙の意味を考慮していないために発生しているといえる．しかし，深層的に見ると，Web 全体の仕組みや自然言語処理技術の現状に大きく依存した問題となっており，この完全なる解消は相当困難である．

一方，1998 年に，Web の創始者である Tim Berners-Lee によってセマンティック Web[2]という概念が提案されており，現在では，このセマンティック Web によるトピックドリフト問題の根本的解決が期待されている．セマンティック Web では，メタデータと呼ばれるデータを説明するためのデータを用い，Web ページの意味内容の公開を行う．そして，検索要求を持つ利用者はエージェントに検索要求を与え，エージェントは与えられた検索要求を解析し検索語を決定した後，メタデータの探索を行う．そして，発見したメタデータに対して，RDF（Resource Description Framework）スキーマ[3]とオントロジ[4]を用いてメタデータを分析する．この際，検索要求に対する意味理解を行い，メタデータを解釈する．このようにして，検索要求に見合った Web ページを発見することができる．

上述のように，セマンティック Web は非常に強力かつ有用な概念である．しかし，Tim Berners-Lee は概念モデルの提案にとどまっており，実際のメタデータへのアクセス方法については敢えて主張をしていない．

こういった状況の中，我々は，セマンティック Web を実現するために，このメタデータへのアクセス方法について今後研究が必要であると考えている．そして，この考えに基づき，効率的なメタデータ収集法の提案を行ってきた[5][6]．本論文では，我々がこれまで提案してきた効率的なメタデータ収集法の評価について報告する．

†岩手県立大学大学院 ソフトウェア情報学研究科

2. メタデータとその収集に関する考察

セマンティック Web で用いられるメタデータは，RDF ファイルとして公開される．RDF とは，W3C[7]で提案されたメタデータを記述するための枠組みであり，記述言語としては XML（eXtensible Markup Language）を用いる．RDF では，リソース，プロパティ及び値の 3 つの項目を用い，リソースにおける値の意味をプロパティで示し，ステートメントとして記述する．RDF のモデルは項目と項目の関係を簡潔に表現するものであり，否定的な表現や，曖昧な表現が含まれておらず，明確に情報の意味を記述することができる．ソフトウェアは，この RDF に従い記述されたメタデータを参照することによって，Web ページの詳細な意味内容を知ることができ，その結果，Web ページの理解が可能となる．

2.1 既存のメタデータ収集法

既存のメタデータ検索システムとして，Weblog に特化したメタデータ検索システムが存在している．これら既存のメタデータ検索システムにおけるメタデータ収集法は，Web ページに付与されているメタデータを人が手動登録する方法と，自動収集する方法の 2 種類に大別できる．

このうち手動登録する方法は，実現の容易さなどの利点があるものの，いずれ限界になると考える．なぜなら，メタデータの増加に伴い，第 3 者によるメタデータの登録が困難となるほか，作成者による登録においても，管理するメタデータが増えた場合煩わしくなるからである．また，もう一方の自動収集する方法は，Weblog の仕組みに大きく依存しているため，一般の Web ページに付与されたメタデータを収集できない．従って，本研究では手動登録ではなく，自動収集を中心に，一般の Web ページに付与されたメタデータを収集する方法を対象としている．

2.2 Web ページ収集法の応用と限界

メタデータを自動収集する方法として，既存の Web ページ収集法の応用が考えられる．しかし，この Web ページ収集法を用いて同様にメタデータを収集しようとした場合，以下の 2 点が問題となる．

(1) Web ページから自身のメタデータの位置を特定することが困難
(2) メタデータとメタデータは互いにリンクを張るようなものではないため，あるメタデータを起点として収集することが不可能

問題点（1），（2）は，どちらもメタデータ本来の用途に関連したものである．メタデータとは Web ページの意味内容を記述するものであるため，メタデータから対象の Web ページの位置を特定できる必要はあるが，Web ページから自身のメタデータの位置を特定できる必要は一般的にはない．従って，（1）に示すように，Web ページからメタデータの位置を特定し，収集することは困難である．また，メタデータは特定の Web ページの意味内容を記述す

るものであり，他のメタデータへリンクを張るようなものではない．従って，（2）に示すように，あるメタデータを起点として収集しようとしても，他のメタデータを発見することができない．

3. 効率的なメタデータ収集法の提案

3.1 メタデータの実際の利用状況の分析調査

我々は，効率的なメタデータ収集法を考案するために，メタデータの利用状況について調査を行った．

調査方法としては，Google のファイルタイプ検索を用い，検索語を「filetype:rdf RDF」にして検索後，検索結果を分析する方法をとった．

Google を使用した検索によって適合する RDF ファイルは，Web ページからリンクされているものになる．本来ならば，リンクされていないメタデータも含めて調査すべきであるが，メタデータは一般的に Web ページからリンクされておらず，リンクされていないメタデータを対象に検索を行う方法は現在知られていない．このため，Web ページからリンクされているメタデータを対象として調査を行った．本調査では，検索結果 33,400 件のうち，上位 100 件に対する分析調査を行った．検索結果の上位 100 件に対象を絞った理由は，Google の検索結果の上位にあるメタデータは PageRank の原理によって良く参照されており，ある程度意味のあるメタデータであると考えたためである．本分析結果を**表 1** に示す．

表 1 の分析結果から，「index.rdf」，「ドメイン名の一部.rdf」，「RDF ファイルが存在するフォルダ名.rdf」といった名称での利用が多いことがわかった．続いて，分析対象を増加させた場合の変化を見るために，分析対象を 300 件に増加させ，同様の分析調査を行った．本分析結果を**表 2** に示す．

表 2 より，分析対象を増加させても，分析対象が 100 件であった場合と同様であり，主要部分の割合の変化はほとんど見られないことがわかる．ただし，100 件を対象にした場合には見られなかった，「diary.rdf」，「foaf.rdf」，「rss.rdf」といった名称での利用を確認した．

次に，時間の推移による変化を調べるために，分析対象を上位 50 件にして，3 ヶ月間の推移を調査した．本分析結果を**表 3** に示す．

表 3 の分析結果から，時間の経過によって主要部分の割合がほとんど変化していないことがわかる．また，「weblog.rdf」といった名称での利用を確認した．

3.2 メタデータの設置場所ガイド

メタデータの調査を行っている過程で，2004 年 5 月に INTAP セマンティック Web 委員会[8]からメタデータの設置場所ガイドが提案された．この設置場所ガイドはメタデータの設置方法について説明し，どのようにメタデータを設置するのが妥当かを提案している．提案されている設置方法は，以下の 3 つである．

(i) コンテンツごとにメタデータをコンテンツに埋め込まずに記述する方法
(ii) 複数のコンテンツに対し 1 つのメタデータをコンテンツに埋め込まずに記述する方法
(iii) コンテンツ内にメタデータを埋め込む方法

このメタデータ設置場所ガイドは，メタデータをどこにどのように設置するかを提案しているものであり，どのような名前をつけるべきかということは規定していない．そのため，メタデータ作成者はどのようなファイル名をつけても良いことになっている．しかし，メタデータを収集する立場からは，どのようなファイル名でメタデータを設置しているかを知る必要がある．従って，我々は，メタデータの利用状況の分析調査によって発見した結果を用いて，メタデータがどのようなファイル名で置かれているかを推察しながら，メタデータ設置場所ガイドに基づいて設置場所及び設置方法を特定する効率的なメタデータ収集法の提案を行っている．

表 1 検索結果の上位 100 件についての分析結果

ファイル名	上位 100 件中の割合
index.rdf	58.0%
ドメイン名の一部.rdf	6.0%
RDF ファイルが存在するフォルダ名.rdf	3.0%
その他（特に特徴のなかったもの）	33.0%

表 2 検索結果の上位 300 件についての分析結果

ファイル名	上位 300 件中の割合
index.rdf	62.3%
ドメイン名の一部.rdf	3.0%
RDF ファイルが存在するフォルダ名.rdf	1.7%
diary.rdf	1.3%
foaf.rdf	5.0%
rss.rdf	4.3%
その他	22.0%

表 3 検索結果の上位 50 件の推移についての分析結果

ファイル名	2004/9	2004/10	2004/11	平均
index.rdf	52.0%	56.0%	58.0%	55.0%
ドメイン名の一部.rdf	8.0%	4.0%	8.0%	6.0%
RDF ファイルが存在するフォルダ名.rdf	2.0%	2.0%	0%	1.3%
foaf.rdf	14.0%	8.0%	6.0%	9.3%
rss.rdf	4.0%	8.0%	12.0%	8.0%
weblog.rdf	4.0%	4.0%	4.0%	4.0%
その他	16.0%	24.0%	12.0%	17.0%

3.3 効率的なメタデータ収集法[5][6]

我々の提案している効率的なメタデータ収集法は，Web ページ収集法を応用した単純なメタデータ収集法[9]と，利用状況に基づいたヒューリスティクスから構成される．以下に本効率的なメタデータ収集法の詳細を示す．

単純なメタデータ収集法[9]

(a) Web ページ内に，RDF ファイルへのリンクがある場合に収集
(b) ディレクトリリスティングが許可されている場合，ディレクトリから RDF ファイルを収集
(c) Web ページの URL の拡張子を「.rdf」に変更し，そのようなファイルが存在した場合に収集

利用状況に基づいたヒューリスティクス

(d) 「index.rdf」が存在する場合に収集
(e) 「ドメイン名の一部.rdf」が存在する場合に収集
(f) 「Web クローラが取得した Web ページの存在するフォルダ名.rdf」が存在する場合に収集
(g) 特定の名前（「diary.rdf」，「foaf.rdf」，「rss.rdf」，「weblog.rdf」）が存在する場合に収集

4.効率的なメタデータ収集法の評価

4.1 評価目的

本研究では，メタデータを効率的に収集することを目的としている．従って本評価では，メタデータを効率よく収集できているかといった観点から，基本性能を示すことが必要である．また，我々の提案している収集法では，効率を重視して収集を行っているため，収集したメタデータが，利用者にとってどの程度有用であるかといったことも示す必要がある．従って，以下の項目について評価を行う．

- メタデータの収集数
  効率的なメタデータ収集法を実装したメタクローラが，どれだけ多くのメタデータを収集できるかを計測
- 収集速度
  効率的なメタデータ収集法を実装したメタクローラが，何時間で何件のメタデータを収集できるかを計測
- 収集したメタデータの有用性
  効率的なメタデータ収集法によって収集したメタデータが，どの程度，利用者の満足度を満たすことができるかを調査

4.2 評価方法

以下に，メタデータの収集数，収集速度，収集したメタデータの有用性についての評価方法を示す．

- メタデータの収集数

単純なメタデータ収集法を用いた単純なメタクローラと，効率的なメタデータ収集法を用いた効率的なメタクローラを実装し，その収集数を比較する．

両メタクローラの動作概要を以下に示す．ただし，最初に，URL を格納する URL データベースに初期 URL リストを与えておくものとする．

(1) URL データベースに接続を行い，URL リストを取得する．そして，URL リストに含まれる URL の全てに対して取得要求を行い，Web ページが存在する場合，その Web ページを収集する
(2) 収集した Web ページを参照し，この Web ページからリンクされている Web ページの URL が存在する場合，URL データベースに追加する
(3) URL リストに含まれる全ての URL に対し，単純なメタデータ収集法または効率的なメタデータ収集法に従って探索を行う．そして，メタデータが存在する場合，メタデータを収集し，メタデータデータベースに追加する

ただし，最初に格納する初期 URL リストとして，100 件の URL を利用することとし，リンクを辿ることによってメタデータを収集できる可能性がある URL 群（リスト α），メタデータが付与されている可能性がある URL 群（リスト β），メタデータが付与されている可能性がありかつリンクを辿ることによってメタデータを収集できる可能性がある URL 群（リスト γ）の 3 種類を用意する．こ

表 4 初期 URL リストの一部

分類	URL
リスト α	・www.yahoo.co.jp ・www.livedoor.com ・www.infoseek.co.jp ・www.msn.co.jp
リスト β	・www.asahi.com ・www.japan.cnet.com ・www.itmedia.co.jp/news ・www.watch.impress.co.jp
リスト γ	・daml.semanticweb.org ・www.kanzaki.com/docs/sw ・www.w3.org/2001/sw ・semblog.org

のような初期 URL リストの一部を表 **4** に示す．この 3 種類の初期 URL リストに対し，それぞれ収集数を計測する．

- 収集速度

前述のメタクローラの起動から終了までの時間を計測し，単位時間当たりの収集速度を算出する．そしてその速度の比較を行う．

- 収集したメタデータの有用性

収集したメタデータの有用性検証のため，検索サービスの実装を行う．そして，この検索サービスを用いて，効率的なメタデータ収集法に従い収集したメタデータが，どの程度利用者の検索要求を満足するかをアンケートによって調査する．本検索サービスは，以下のメタクローラ，メタデータデータベース，メタデータリトリーバから構成される．

メタクローラ

メタクローラは，効率的なメタデータ収集法に従い，メタデータを収集後，メタデータベースに蓄積する．

メタデータデータベース

メタデータデータベースは，収集したメタデータを蓄積するデータベースである．

メタデータリトリーバ

メタデータリトリーバは本検索サービスのユーザインタフェースであり，利用者からの検索要求を検索語として受け付ける．また，メタデータリトリーバは，検索語を取得後，メタデータデータベースに含まれるメタデータに対して全文検索を行い，検索語と一致するメタデータのタイトル，URL，概要を検索結果として提示する．

4.3 評価結果

評価方法に従い評価を実施した．以下に，評価結果を示す．

- メタデータの収集数

単純なメタクローラ及び効率的なメタクローラの収集数の計測結果をそれぞれ表 **5**，表 **6** に示す．

表 **5**，表 **6** の収集数の結果から，収集数が平均で約 **4.4** 倍になっていることがわかった．また，効率的なメタデータ収集法を構成する各方法ごとの収集数を表 **7** に示す．

表 **7** に示すように，リスト α 及びリスト β では，方法 (d) 以外での収集を確認できなかった．これに対し，リスト γ では，すべての方法での収集を確認できた．全体的に見て，方法 (e)，(f)，(g) での収集数が少ないこと

表 5 単純なメタクローラの収集数の計測結果

分類	対象となった URL 数	収集数
リスト α	4063	2
リスト β	3266	0
リスト γ	5876	110
平均		37.3

表 6 効率的なメタクローラの収集数の計測結果

分類	対象となった URL 数	収集数
リスト α	4063	85
リスト β	3266	49
リスト γ	5876	373
平均		169.0

表 7 各方法ごとの収集数の計測結果

分類	収集数	単純なメタデータ収集法	方法 (d)	方法 (e)	方法 (f)	方法 (g)
α	85	0	85	0	0	0
β	49	0	49	0	0	0
γ	373	91	261	6	5	9

がわかった．

● 収集速度

単純なメタクローラと効率的なメタクローラの収集時間に関する計測結果を**表 8** に示す．

表 8 の結果から，単純なメタクローラの収集時間は平均で約 314.7 分であり，効率的なメタクローラの収集時間は平均で約 331.9 分であった．メタデータの収集数が平均で約 4.4 倍となるのに対し，収集時間は約 5.4%増にとどまっている．収集時間と収集数から，メタデータを 1 件収集するのに必要な収集時間を算出すると，単純なメタデータ収集法のメタデータ 1 つあたりの収集時間は，平均で約 33.6 分であり，効率的なメタデータ収集法のメタデータ 1 つあたりの収集時間は平均で約 2.3 分である．また，収集速度については，単純なメタクローラは平均で約 7.3（件/時）であるのに対し，効率的なメタクローラは平均で約 30.6（件/時）である．

このことから，効率的なメタデータ収集法は，単純なメタデータ収集法に比べ，メタデータを効率的に収集できることがわかる．

● 収集したメタデータの有用性

収集したメタデータの有用性に関する調査結果を**表 9** に示す．**表 9** では，4 人の被験者を被験者 A，被験者 B，被験者 C，被験者 D と表している．被験者には，興味のある分野について任意の検索語を選定させ，1 人あたり 3 回の検索を実施させた．その結果，全ての被験者が 3 回中少なくとも 1 回は検索要求に一致する Web ページを発見することができた．しかし，一部の検索語に対しては，検索要求と一致する Web ページを発見することができない場合もあった．この原因は，メタデータベースに蓄積されているメタデータの総数が少ないことであると考えられる．効率的なメタデータ収集法によって，従来よりも多くメタデータを収集することが可能となったが，本評価では 500 件程度のメタデータに対し検索を行っているため，常に被験者

表 8 両メタクローラの収集時間

分類	単純なメタクローラの収集時間	効率的なメタクローラの収集時間
リスト α	121 分 57 秒	141 分 50 秒
リスト β	135 分 26 秒	159 分 9 秒
リスト γ	686 分 40 秒	694 分 40 秒
平均	約 314.7 分	約 331.9 分

表 9 収集したメタデータの有用性の調査結果

	検索要求	検索語	Hit 数	満足度
被験者 A	オートバイ	カブ	2	不満
	映画	ハウル	2	満足
	ユビキタス	ユビキタス	4	満足
被験者 B	ゲーム	テトリス	1	満足
	Java 言語	Java	5	不満
	IT 技術	IT	3	満足
被験者 C	HTML	HTML	3	不満
	Linux	Linux	3	満足
	Macintosh	Mac	2	不満

の与えた検索語と一致させるのは困難であったと考える．

5.おわりに

本論文では，我々がこれまで提案してきたセマンティック Web における効率的なメタデータ収集法の評価について報告を行った．本評価では，効率的なメタデータ収集法によってメタデータを効率的に収集することが可能であることがわかった．

今後の課題として，利用状況の継続調査や，利用状況が変化した場合の効率的なメタデータ収集法の拡張などが挙げられる．

参考文献

[1]荒谷 寛和，藤田茂，菅原研次：エージェントに基づくウェブページ分類の実験評価（1），電子情報通信学会技術研究報告，Vol.103, No.243, pp.49--54 (2003).

[2]http://www.w3.org/2002/Talks/04-sweb/

[3]松井 くにお，津田 宏，上田 健次，小泉 雄介，豊内 順一，布目 光生：セマンティック Web におけるメタデータとその活用，情報処理，Vol.43, No.7, pp.718--726 (2002).

[4]清野 正樹，来間 啓伸，今村 誠：セマンティック Web とオントロジ記述言語，情報処理，Vol.43, No.7, pp.727--733 (2002).

[5]及川啓，児玉英一郎，高田豊雄：セマンティック Web における効率的なメタデータ収集法に関する考察，平成 16 年度電気関係学会東北支部連合大会講演論文集，p.228 (2004).

[6]及川啓，児玉英一郎，王家宏，高田豊雄：セマンティック Web における効率的なメタデータ収集法の提案，マルチメディア，分散，協調とモバイルワークショップ論文集，(2005, to appear).

[7]http://www.w3.org/

[8]http://www.net.intap.or.jp/INTAP/s-web/index.html

[9]清水佳奈，佐々木亮，児玉英一郎，高田豊雄：Web を利用した次世代の楽譜利用環境に対する考察，平成 15 年度電気関係学会東北支部連合大会講演論文集，p.387 (2003).