E-004

Web ページに対する RDF メタデータ付与支援ツール開発
Development of a tool for assigning RDF metadata to Web pages

佐藤 邦俊† 桂田 浩一† 山田 博文‡ 新田 恒雄†
Kunitoshi Sato Kouichi Katsurada Hirobumi Yamada Tsuneo Nitta

1. はじめに

機械可読な RDF メタデータ情報を Web ページに付加することで，様々な情報提供や問題解決を行う Semantic Web 技術の研究が盛んになっている[1]．しかしインターネット上に氾濫する大量の Web ページに RDF を手動で付与することは，ユーザにとって大きな負担であり Semantic Web 普及の妨げの一因となっている．我々はこの問題を解決するために，ブラウザ上での簡単な操作によって RDF メタデータを作成する機能，および RDF メタデータを半自動的に付与する機能を持つ支援ツールの開発を目指している．本報告ではツールの機能のうち，RDF 手動付与の機能と，半自動付与の前処理である Web ページからの重要語自動抽出法について説明し，関連研究と比較する．

2. RDF メタデータ付与支援ツールの概要

RDF は Web 上にある様々な「リソース」（Web ページやその一部など，何でも良い）に，その「プロパティ」と「値」の組（例えば「ページの作者」は「佐藤」である等）を与える書式の一つである．今回開発した本ツールは，ユーザがブラウザ上からマウス操作で Web ページの一部を選択し，それを「リソース」もしくは「値」として RDF を付与する機能と，Web ページ内の重要語を「値」の候補として自動的に取得し，RDF を付与する機能を提供する．以下に各機能について説明する．

2．1 マウスによる RDF 付与機能

マウス操作による RDF 付与では，ブラウザに表示されている Web ページから，そのページを表現していると思われる特徴的な単語やフレーズをユーザが，マウスで選択した後，「プロパティ」を指定する．このとき、ユーザは選択した単語，フレーズを「値」もしくは「リソース」として利用できる．「値」として用いた場合，RDF の「リソース」は Web ページの URL，または，選択した語やフレーズの位置(Xpointer)になり，Web ページに意味付けをしたとみなされる．一方，選択した単語やフレーズを RDF の「リソース」として用いる場合は，選択した単語やフレーズに別途「値」を与えて RDF を作成することになる．

開発したツールは RDF 付与のインタフェースとして，付与ダイアログを用いる方法と，メニューから簡単に付与する方法の二つを提供している．前者は，「リソース」を示す「値」に単語やフレーズ，他のメタデータを付与する等，細かい設定ができるのに対して，後者は，選択した単語かフレーズのみを利用できるだけの簡易機能である．

†：豊橋技術科学大学　大学院工学研究科
‡：豊橋技術科学大学　マルチメディアセンター

2．2 重要語自動抽出による RDF 付与支援機能

2.2.1 機能の概要

対象の Web ページを表現していると思われる特徴的な単語やフレーズを自動的に抽出し，ユーザに提示することで，RDF 付与を支援する．この機能を使うことによって，Web ページ内の特徴的な語をユーザ自身が発見する負担が減るため，RDF 作成は更に簡単になる．将来はこの機能に，プロパティ付与の傾向を学習する機能を付け，RDF メタデータの半自動生成を行う予定である．なお，重要語の自動抽出による RDF メタデータ作成では，リソースは Web ページの URL となる．

2.2.2 重要語自動抽出アルゴリズム

本ツールの重要語自動抽出は，大規模コーパスを対象とする訳ではなく，Web ページ，もしくは，そのページと同じドメイン下の Web ページ群を対象にしている．Web ページに対する重要語抽出では，類似文書を大量に必要とするキーワード抽出法を採用したり，多様な Web ページから適切なコーパスを収集する方法を適用することは困難である．そこで，本ツールでは，語の共起関係を基にして統計的指標を用いる方法を採用した[2]．この方法は，少ない文書からでも重要語を抽出することができる．採用した手法は，Web ページ内のある語とそのページの頻出語群との共起確率，および頻出語単独の生起確率間の分布の偏りを調べ，ずれの大きな語を重要語として抽出する．この時，統計的に有意なずれを評価するため，χ^2 検定を用いて分布の偏りを検定している．手法の詳細は文献[2]を参照されたい．

2．3 その他の機能

RDF メタデータの作成では，すでに定義されているプロパティ以外のユーザ定義プロパティを利用する場合，プロパティを規定する RDF Schema を作成する必要がある．本ツールは，プロパティを定義する機能を持っており，それを使用することによって RDF Schema を自動的に作成できる．なお，プロパティを定義する際には，RDF メタデータ作成と同様に，必要情報を選択するだけで容易に作成できるようになっている．

2．4 ツールの実行例

2.4.1 マウス操作によるＲＤＦ付与

ブラウザに表示された Web ページから，ユーザがそのページを表現していると思われる特徴的な単語やフレーズをマウスで選択(図１左)し，メニューから「ＲＤＦ付与ダイアログを開く」か「簡易的にＲＤＦを付与」のどちらかを選択する．前者を指定すると，RDF 付与ダイアロ

図１．マウスで語の選択(左)と RDF 付与のマーク表示(右)

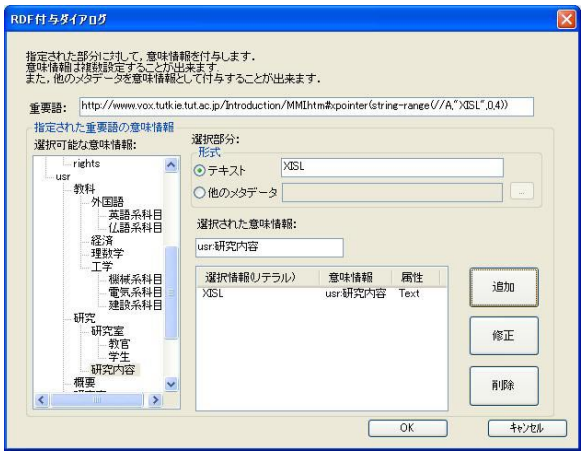

図２．RDF の付与画面

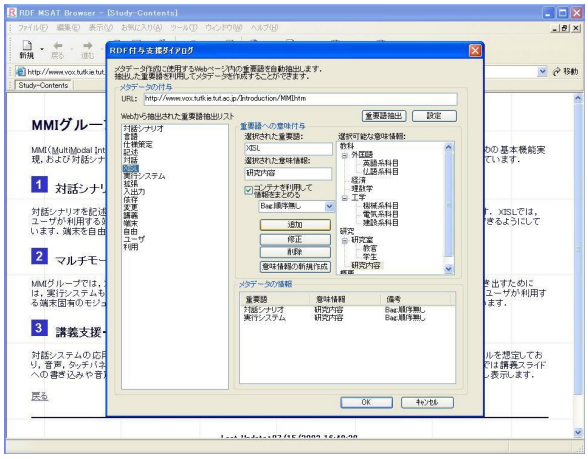

図３．重要語の自動抽出による RDF 付与画面

グが表示されるので，プロパティリストから付与したい項目を指定して追加を押すと，RDF が付与される(図２)．後者を選ぶと，メニューに現在登録済みのプロパティが表示されるので，その中から使用したいプロパティを選択することで，RDF を簡単に付与できる．なお，RDF 付与を終了すると付与を示すマークが表示される(図１右)．このマークをダブルクリックすることで，編集ダイアログを開くことができる．

2.4.2　重要語自動抽出によるＲＤＦ付与

ツールバーから「重要語抽出」を選択すると，RDF 付与支援ダイアログ(図３)が表示される．続いて重要語抽出のボタンを押すと，指定された URL の Web ページを解析し，ページ内の特徴的な語を重要語として抽出する．重要語の抽出後は，ユーザが重要語を選択し，使用したいプロパティを選択することで RDF を付与できる．

3．関連研究との比較

先行研究として様々なツールが提案されている．以下では，本研究と同様にメタデータ付与支援を目的としたツールや，他の重要語抽出手法を本ツールと比較する．

3．1　メタデータ支援ツール DC-dot と比較

DC-dot は，書誌情報の特徴を統一的に記述する RDF スキーマの一つである，Dublin Core を使用したメタデータを自動生成する支援ツールである[3]．Web 上の入力ファームから対象となる Web ページの URL を入力することで，ページから情報を抽出し Dublin Core メタデータを作成する．本ツールとの相違点は，DC-dot がプロパティを Dublin Core で規定されたもののみとしているのに対して，本ツールではリソースやプロパティをユーザが自由に選択，定義できるようにしている点である．本ツールでは Dublin Core などの既存のプロパティだけでなく，ユーザが規定したプロパティも使用して RDF メタデータの作成を行うことができる．

3．2　注釈付加ツール Annotea との比較

Annotea は Web ドキュメントに注釈付けを行うツールである[4]．注釈はメタデータとしてモデル化され RDF 形式で保持される．作成された注釈は Web 上の共有サーバに保存され，複数の利用者が注釈を共有することができる．一方，本ツールはユーザ嗜好の RDF メタデータを作成することを目標とするため，データの共有化は行っていない．しかし，メタデータの半自動付与の際には，多くの RDF メタデータが学習データとして必要になる．他のユーザのメタデータを利用することも検討する必要があると考えている．

3．3　固有表現抽出による重要語抽出との比較

固有表現抽出は，文書中の「豊橋の佐藤」等の固有表現に<ADDRESS>や<PERSON>などの意味タグを付与する技術である．意味タグを付与した文書から，指定したタグ要素を重要語として抽出する[5]．この手法は，例えば，他社製品名を抽出して，その製品に関する情報を抽出・収集してユ提供する等のアプリケーションに有効である．本ツールは，特定のアプリケーションに限定せず，様々な Web サイトを対象に，メタデータを作成することを目的としている．このため，固有表現抽出技術を使用せず，対象の Web ページを表現している特徴的な語を重要語として抽出する手法を用いている．

4．ま と め

本報告では，開発中の RDF メタデータ付与支援ツールについて述べた．マウス操作による RDF 付与や重要語自動抽出による付与支援機能は，ユーザの負担を軽減し，RDF メタデータの作成を容易にする．今後は，重要語抽出精度を向上させると共に，抽出した重要語に対してユーザが作成したメタデータを用いて，プロパティ付与傾向を学習したり，タグ構造，シソーラス等を利用して RDF メタデータを自動生成する機能を追加していく予定である．

参考文献

[1] 荻野 達也他：“セマンティック Web とは”，情報処理学会誌，Vol.43，No.7，pp.709-717 (2002).
[2] 松尾 豊他：“語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム”，人工知能学会論文誌，Vol.17，No.3，pp217-223(2002).
[3]http://www.ukoln.ac.uk/metadata/dcdot/.
[4]Kahan，J.et al.：Annotea：An Open RDF Infrastructre for Shared Web Annotations，Proc. The 10th International Conference on World Wide Web， ACM Press， pp.623-632(2001).
[5]松平 正樹他：“文書からのキーワード抽出と関連情報の収集”，人工知能学会研究会資料，SIG-SWEO-A3-03-02，pp02－01 - 02－06(2004).