

Web ページの重要語を利用した RDF メタデータ付与支援ツールの開発*

佐藤 邦俊† 桂田 浩一† 新田 恒雄†
豊橋技術科学大学 大学院工学研究科 †

1. はじめに

近年, 新しい Web 技術として, Semantic Web の研究が盛んになっている[1]. Semantic Web は, RDF という機械可読なメタデータを Web ページに付加し, 情報提供や問題解決を行うための枠組みである. しかし, このメタデータを作成することが, Semantic Web 普及の妨げる問題の一因となっている. 原因として, インターネット上の大量の Web ページに付与しなければならないことが挙げられる. これを手動で行うことは, メタデータ作成者に大きな負担となる.

我々は, この負担の軽減を目的として, Web ページの複合語を重要語として提示し, 簡単な操作で RDF を作成できるツールを開発している. 本報告ではツールの概要と, Web ページから重要語を自動抽出する手法について述べる.

2. RDF メタデータ付与支援ツールの概要

RDF は, リソースと呼ばれる Web ページやその一部に「プロパティ」と「リテラル」の組を与えて作成する. 「ページ A の作成者は佐藤である」を例にとると, ページ A がリソース, 作成者がプロパティ, 佐藤がリテラルになる. これまで, この RDF 作成の支援を行うために, 自動付与やデータ共有などを行うツールが開発されている[2][3]. 今回開発した支援ツールは, メタデータ作成者に Web ページ内にある重要語を提示することで支援を行う.

2.1 RDF の付与

RDF 付与には, 図1に示す RDF 付与ダイアログを使用するか, あるいは簡易付与機能を使用する. 前者は, RDF メタデータを作成する際に使用する標準機能である. リテラルにテキストだけでなく他のメタデータを付与するなど, 細かい設定ができる. 一方後者は, Web ページ上の字句をリソースもしくは, リテラルとして付与する機能である.

メタデータ作成者は, まず, ツールのメニューから RDF 付与支援ボタンを選択する. すると, Web ページからそのページの上位 20 個の重要語が抽出され, 重要語の背景色が赤色で表示される(図2). その後, 表示されている重要語を参考にして, RDF 付与ダイアログか簡易機能を使用して RDF の作成を行う. 作成が完了すると, 付与したことを示すマークが重要語の前につけられる(図3).

RDF を付与する場合, 通常は, リソースとしてそのページの URI が指定される. 一方, ページ上の字句をリソースとする場合は, その位置を Xpointer で表現する.

プロパティについては, 既存の RDF スキーマに記述されるプロパティを使用するか, 本ツールのプロパティ編集

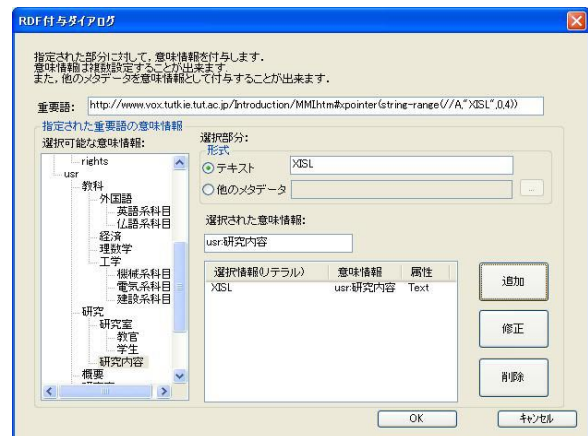


図1. RDF の付与画面

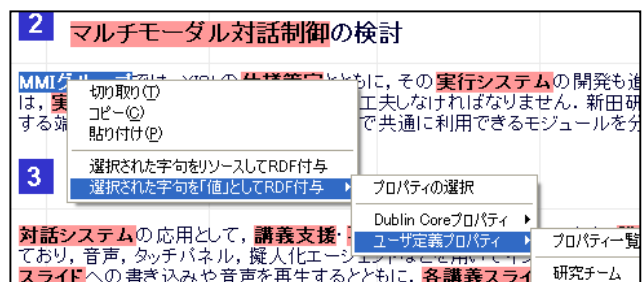


図2. 重要語の表示と字句の選択

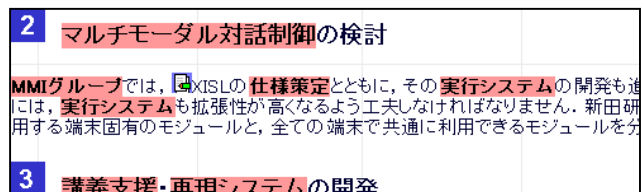


図3. RDF 付与のマーク表示

機能を用いて, ユーザ独自のものを作成して利用する.

2.2 重要語抽出アルゴリズム

2.2.1 重要語抽出アルゴリズム

本ツールでは, 複合語の出現頻度 (TF 値) を主な指標として重要語を抽出する. しかし, Web ページは, 文の数が少ない場合が多く, 重要な複合語であっても必ずしも頻度が多いとは限らない. そこで, 本手法では, 複合語の頻度に加えて, 複合語を構成する各形態素の頻度も利用した.

また, Web ページは, <H1>や等の強調タグや箇条書きタグで, ページの重要な部分を表現することが多い. そこで, 各タグの重要語表現への「寄与度」を予備実験により求め, 数値の高いタグに囲まれている複合語の重要性を高く評価することにした. この「寄与度」は, タ

*A support tool for attaching RDF-metadata to a web page using keywords the page
K.Sato†, K.Katsurada†, T.Nitta†
Graduate School of Engineering, Toyohashi Univ.of Tech. †

グに囲まれている文に、どれだけ重要語が含まれているかを値にしたものである。

一方、予備実験の結果、形態素数が多い複合語は重要である確率が高いという傾向が見られた。そこで、形態素数に基づいたバイアス B_n (複合語 n の形態素数による重要語になりやすさ) を用いることにした。

以上の要素を考慮し、重要語判定に次式を適用する。式中のパラメータ α , β , γ , θ は、それぞれ $\alpha=0.7$, $\beta=0.01$, $\gamma=10$, $\theta=0.5$ とした。

$$I(n,d) = B_n \{ \alpha TF(n,d) + (1-\alpha) TF_{avr}(n,d) \} \left(\frac{1}{1 + e^{-\gamma(C-\theta)}} + \beta \right)$$

$I(n,d)$: Web 文書 d における複合語 n の重要度

$TF(n,d)$: Web 文書 d における複合語 n の出現頻度

$TF_{avr}(n,d)$: Web 文書 d において複合語 n に含まれる各形態素の平均出現頻度

C : 複合語 n または n が含む形態素を括弧しているタグの寄与度の最大値

α : 重み定数

γ : 制御パラメータ(シグモイド関数の傾き)

θ : 閾値パラメータ

β : フロアリング定数

2.2.3 抽出実験

2.2.1 に説明した手法を用いて、Web ページから重要語を抽出する実験を行った。今回は、国内の大学研究室を対象に、研究紹介のページを無作為に 10 ページ選び、被験者 5 名がそのページの重要語を抽出した。次に、過半数の被験者が重要語として抽出したものを正解データとし、上記手法を用いて重要語抽出の評価実験を行った。実験では、重要度の高い複合語上位 40 個まで抽出し、10 個単位で正解データとの適合率、再現率を求めた。

また、比較評価のために、以下の 3 種類の手法で同様の実験を行った。

1. TFIDF 法: 語の出現頻度 (TF 値) と偏り (IDF 値) に基づいて重要語を抽出する。
2. 手法 A: 少ない文から重要語を抽出するために、語の共起関係に基づいた統計的指標を用いる[4]。
3. 手法 B: 複合語の TF 値、複合語の部分列の TF 値、品詞に対する重み、複合語の長さ、複合語の IDF を用いて重要語抽出を行う[5]。

なお、手法 B のパラメータの一つである複合語の部分文字列に対するバイアス係数は、文献[5]では 0.6 としているが、本実験では最も結果の良かった 0.1 に変更した。

2.2.4 結果と考察

実験結果を図 4 に示す。グラフの各点は、左から順に各手法で出力された上位 10 個、20 個、30 個、40 個の重要語の再現率、適合率を表す。図に示すとおり、本手法は他の手法に比べて良好な結果を得た。本研究では、2.1 節で述べたように、メタデータ付与支援ツールで 20 個程度の重要語を抽出することを目的としている。図から、上位 20 個での重要語の再現率、適合率は、手法 B と比較してそれぞれ 12.8%, 11.1%, TFIDF 法とは 18.3%, 17.2%, 手法 A では 30.2%, 26.4% の性能向上が見られた。

性能向上の最大の要因は、他の手法で用いられていないタグの寄与度を利用しているからであると考えられる。実際に実験結果を解析すると、タグの寄与度の利用を除けば、本手法に最も近い手法 B と比較したとき、上位 20 個の重要語につき新たに平均 3.0 個の重要語を検出できてい

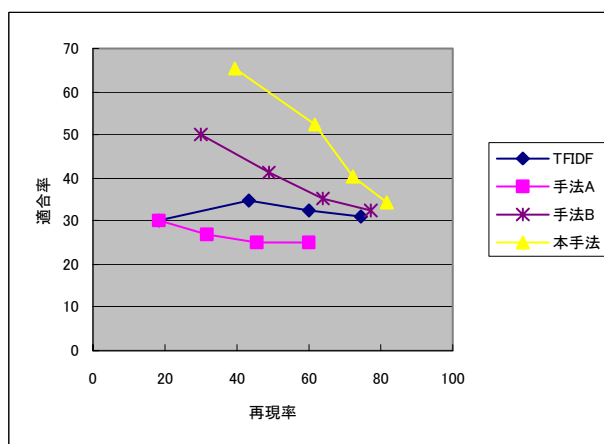


図 4. 各手法の再現率—適合率

る。特に寄与度の高いタグを多く利用しているページでは最大 6 個の重要語を検出できた。一方、寄与度の低いタグが多いページでは、再現率、適合率が変わらない場面もあった。

次に、本手法と手法 B は、TFIDF 法と手法 A と比較したとき、良好な結果を得た。これは本手法と手法 B で利用している複合語の形態素数に対するバイアスが抽出に効果を発揮したためである。実験対象とした研究紹介ページでは、形態素数の多い複合語が専門用語として使用されていることが多い。そのためこれらバイアスが有効に働いたといえる。

一方、本手法では TFIDF と手法 B で用いた IDF 値を利用していないが、結果への影響は少なかった。抽出実験の対象とした Web ページ群では、ページ毎の内容がそれぞれ独立している。そのため、複数のページにわたって同じ複合語が出現する確率が低く、単語の偏りが少なかったためである。

3. まとめ

本報告では、RDF メタデータ作成の負担軽減を目的として、Web ページの重要語を利用した RDF 作成支援ツールについて述べた。特に、提案した重要語抽出手法は、HTML タグの寄与度等の利用により TFIDF 法等と比べ、高い抽出精度を達成した。今後は、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] <http://www.ukoln.ac.uk/metadata/dcdot/>.
- [3] Kahan, J. et al.: Annotea: An Open RDF Infrastructure for Shared Web Annotations, Proc. The 10th International Conference on World Wide Web, ACM Press, pp.623-632(2001).
- [4] 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, pp.217-223(2002).
- [5] 西坂 信広他: “日本語文書における複合語キーワード抽出”, 情報処理学会 第 59 回全国大会, Vol.2, pp.321-322. (1999)