Wordlogue: 頻度付き語彙列を用いたブログの分類と検索

長尾　確 †　東中 竜一郎 ††
† 名古屋大学 情報メディア教育センター　††NTT コミュニケーション科学基礎研究所

1　はじめに

ブログ内のテキストへのタグ付けとソーシャルブックマークとの連携により、頻度付き語彙（複合語を含む）列によってブログエントリーの特徴を表現し、その XML 表現を RSS に含めて配信することにより、ブログの分類と検索を支援する仕組みを提案する。このシステムを、Wordlogue（ワードローグ）と呼ぶ。

いわゆるブログ検索サービス（たとえば、テクノラティ[3]）のクローラでも同様なことが可能であると思われるが、それをしない理由は、ブログの著者が自分で確認しながら作成する方がきめ細かな調整が可能で、自分のブログに興味を持たせる手段にもなるからである。

また、頻度は単純な出現頻度だけでなく、照応や省略のアノテーションも考慮して、代名詞や主語等の省略によって参照されている語も重複して出現しているものとみなす。これは、本来重要であるべき語の出現頻度が少ない場合に、重要性を正しく反映させることができる。頻度が同じ語に対して、ソーシャルブックマークに関連する記事が多い語の重要性を高く設定することもできる。

Wordlogue サーバーはブログの更新情報を含む RSS をトランスコードし、頻度付き語彙列の情報を RSS と統合する。それにより、ブログ分類・検索サービスをより高度にすることができる。

2　Wordlogue

一般にブログにはカテゴリーが付与され分類に利用されているが、ブログエントリー内に複数のトピックが含まれることもあるため、単一のカテゴリーに属するには不適切な場合がある。そのため、ブログエントリー内の任意の語（あるいはその上位語）がそのエントリーを代表するキーワードになり得ると考え、テキストを解析して内容語を抽出し、出現頻度の高いものを選択するという手法が考えられる。

しかし、一般に同じ人間が同じ文脈では同じ対象や行為を表すために同じ語を用いることが少ないため、単純な出現頻度では中心的な話題が反映されないことがある。また、通常の形態素解析では、できるだけ詳細に解析するため、必要以上に細かく文を分解することが多いが、たとえば、固有名詞や書名などはより細かく分割すべきではないだろう。つまり、そのような語をまとめて一つのものとして扱う必要があるだろう。

Wordlogue は、ブログオーサーが、ブログエントリーを投稿する際に、まず形態素解析を行い、その後、語の分割・統合、代名詞の照応、用言の主語・目的語の省略に関する言語的アノテーションを付与する仕組みを提供する。その結果は RDF[4] 形式でデータベースに保存され RSS をトランスコードする際に利用される。

図 1: 形態素解析結果の編集画面

3　アノテーションウィザード

Wordlogue で操作するアノテーションは、基本的に 2 つに分けられる。一つは文の分割および内容語の判定で、形態素解析結果を操作し、複合語の統合（たとえば書名や固有名詞などは一語にする）や未知語への品詞の付与を行う。もう一つは、照応や省略に関する、語間のリンクである。今回は複雑な操作を必要とするエディタではなく、より簡易的なものとして、Web 上のウィザードを作成した。これをアノテーションウィザードと呼ぶ。

アノテーションウィザードの利用手順は以下の通りである。

1. Wordlogue にログインし、文章を書く。ブログサーバーの XML-RPC[5] の URL とブログ ID をあらかじめ設定しておく。

2. 次に、ブラウザの解析ボタンをクリックして、形態素解析結果を閲覧する。このとき自動的にソーシャルブックマークが検索され、関連記事の件数が表示される（図 1 を参照）。形態素解析には、奈良先端科学技術大学院大学で開発された茶筌 [6] を用いている。また、ソーシャルブックマークは del.icio.us[1] を用いている。

3. ここで、形態素解析結果を修正する。これは、主に、形態素解析で必要以上に分割された複数形態素を結合することによって行われる。形態素解析結果の修正に応じて、ソーシャルブックマークの検索結果も自動的に変化する。

4. 次に、未知語への品詞付与を行う。これは茶筌の採用している日本語品詞体系をリスト表示して、

Wordlogue: Weblog Classification and Retrieval Using Data on Word Frequency and Annotation

† NAGAO, Katashi(nagao@nuie.nagoya-u.ac.jp)
†† HIGASHINAKA, Ryuichiro(rh@cslab.kecl.ntt.co.jp)
Center for Information Media Studies, Nagoya University (†)
NTT Communication Science Laboratories (††)

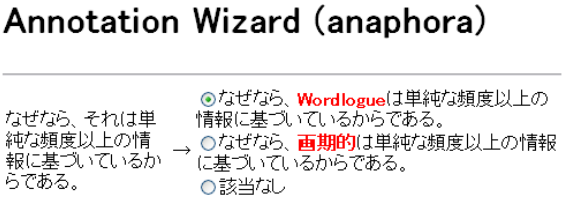

図 2: 代名詞への先行詞情報の付与

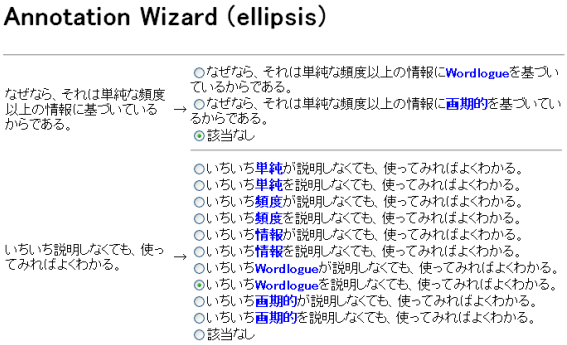

図 3: 省略語の補完情報の付与

適切なものを選択することによって行われる。

5. 次に、代名詞の照応に関するアノテーションを付与する。これは、代名詞をその先行詞（名詞）の候補で置き換えた文を表示して、適切な文を選択することによって行われる（図 2 を参照）。

6. 次に、用言の主語と目的語の省略に関するアノテーションを付与する。これも照応の場合と同様に、省略語（名詞）の候補および助詞を挿入した文を表示し、適切な文を選択する（図 3 を参照）。

7. 最後に、簡易表示された頻度グラフを確認して、内容を確定する。ブログエントリーは図 4 のようになる。ブログサーバーへのエントリーの登録と同時に、すべてのアノテーション情報を RDF 形式で Wordlogue データベースに登録する。

4 ブログの分類と検索

Wordlogue は、ブログエントリーの詳細な言語情報を用いることによって、ブログサービスの高度化を実現する。そのための重要な機能が、RSS トランスコーディングである。

4.1 RSS トランスコーディング

ブログの特徴として、RSS フィードの機能を用いて、ブログの更新を通知し、検索やマイニングを行うことができる点が挙げられる。ただし、RSS に含めることができる情報はタイトルや本文（の一部）や Dublin Core で規定されているメタデータエレメントに限定されている。

われわれは、ブログサーバーに手を加えることなく RSS を拡張し、語や文の構造、およびソーシャルブックマークへのリンク情報に基づいて、ブログの分類や検索を高度化する手法を開発している。それは、Wordlogue サーバーに蓄えられた情報に基づいて、ブログの生成する RSS をトランスコードしてからフィードする仕組みである。このトランスコードされた RSS を拡張 RSS と呼ぶ。

図 4: 作成されたブログエントリー

Wordlogue サーバーは RSS リーダーやブログ検索サービスからリクエストを受け、ブログサーバーから得られた RSS に Wordlogue データが付与されたエントリーへのリンクが含まれている場合に拡張 RSS を生成して送信する。

拡張 RSS を受け取ったサーバーや RSS リーダーは、Wordlogue の情報を用いて、自動的にソーシャルブックマークに登録したり、ブログエントリー間の類似性を計算することなどができる。

5 今後の課題

Wordlogue データは、ブログエントリーに対するアノテーションという位置づけであり、そのアノテーションの再利用を考えるべきである。たとえば、形態素および照応や省略の言語情報を用いたブログの要約、言い換え、翻訳などが考えられる。これには、セマンティックトランスコーディング [2] の技術が利用可能である。

参考文献

[1] del.icio.us, “ソーシャルブックマーク del.icio.us,” http://del.icio.us/, 2005.

[2] Katashi Nagao, Yoshinari Shirai, and Kevin Squire, “Semantic Annotation and Transcoding: Making Web Content More Accessible,” IEEE MultiMedia, Vol.8, No.2, pp.69-81, 2001.

[3] Technorati Japan, “テクノラティ,” http://www.technorati.jp/home.html, 2005.

[4] W3C, “Resource Description Framework (RDF) Model and Syntax Specification,” http://www.w3.org/TR/REC-rdf-syntax/, 2002.

[5] XML-RPC.Com, “XML-RPC Home Page,” http://www.xmlrpc.com/, 2004.

[6] 奈良先端科学技術大学院大学, “形態素解析システム茶筌,” http://chasen.naist.jp/hiki/ChaSen/, 2003.