7C-2

アノテーションの構造化とその処理の提案

阿部 裕行　伊藤 一成　Martin J. DÜRST
青山学院大学理工学部

1　はじめに

情報化社会の今，膨大な量の Web ページやマルチメディアコンテンツが散在している．その中からユーザの要求にあったコンテンツの獲得は困難である．さらに計算機が自動的に，より高度にコンテンツを扱ってくれるような仕組みが必要である．近年，コンテンツの意味や内容に関する特徴をメタ情報として付与し，付与したデータを計算機処理対象とすることで，元のコンテンツ情報を効率よく検索，要約などを行い，高度に扱うアノテーションの研究が注目されている．

人間にとって記述や理解がしやすい自然言語表現により付与するメタ情報をアノテーションと本稿では定義する．ゆえにアノテーションに対して自然言語処理を行うのは有用である．ただし，自然言語処理単独ではなく，データマイニング処理と複合した新しい解析手法を検討するのが望ましいであろう．本稿では，初めにアノテーションの構造化について検討する．さらにそのデータ構造に基づく，処理手法を提案する．

2　アノテーション

2.1　アノテーションの内部データ構造

アノテーションを RDF (Resource Description Framework) [1] で表現する．内容情報を Dublin Core [2] と Dublin Core の拡張語彙を用いて記述し，さらにそのテキストに言語情報として GDA (Global Document Annotation) [3] の付与を行う．これを一つのアノテーションとする．

2.1.1　内容情報

Dublin Core を用いる．Dublin Core では Web ページや文書の書誌データに付与可能な情報として，基本 15 項目を設定している．その中から `dc:creator`, `dc:date`, `dc:title`, 及び `dc:description` を使用する．しかし，アノテーショングラフ構造を考えると，これらの述語だけでは子から親に対してのリンクはたどれるが，その逆はたどれない．これを解決するために，`dc:relation` の拡張語彙の一つである `dcterms:references` を使用する．

2.1.2　言語情報の付与

`dc:description` の目的語として設定するテキストに対し，産業技術総合研究所の橋田が提案する，多言語間に共通の統語・意味などの言語情報記述に関する XML タグセットである GDA を付与する．その他に自然言語処理の前処理として，テキスト内の名詞に出現頻度を表す属性 `tf` を追加する．

2.2　アノテーションのグラフ構造

すべてのアノテーションはその対象を持ち，さらにアノテーションに対するアノテーションも考えると，アノテーションを子とし，その対象を親とする親子関係が定義できる．すると図 1 のように，アノテーション群をグラフ構造とみなせる．例えば，Annotation1 の親は Content で，子は Annotation4 となる．Annotation1, Annotation2 及び Annotation3 は全て同一の Content に対するアノテーションなので，これらのアノテーションは兄弟アノテーションとなる．

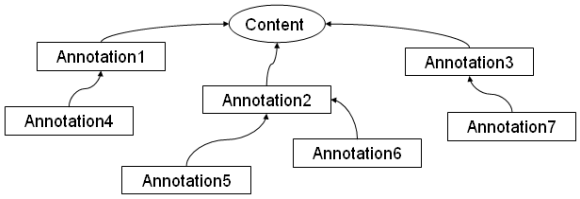

図 1: アノテーショングラフ構造

2.2.1　アノテーション記述例

アノテーション記述例を図 2 に示す．このデータ例は http://www.sw.it.aoyama.ac.jp/2005/hiroyuki/hiroyuki.xhtml に対してのアノテーションである．さらに，このアノテーション自体の URI が http://www.sw.it.aoyama.ac.jp/2005/hiroyuki/hiroyuki.rdf であり，http://www.sw.it.aoyama.ac.jp/2005/hiroyuki/data.rdf からアノテートされていることを示している．

3　自然言語処理

前章で定義した構造化データに対しての言語処理方式について検討する．

A Proposal for Annotation Structuring and Processing
Hiroyuki ABE, Kazunari ITO and Martin J. DÜRST
Department of Integrated Information Technology, College of Science and Engineering, Aoyama Gakuin University
5–10–1 Fuchinobe, Sagamihara, Kanagawa 229-8558, Japan
hiroyuki@sw.it.aoyama.ac.jp, {kaz, duerst}@it.aoyama.ac.jp

```
<rdf:RDF
  xmlns:dcterms="http://purl.org/dc/terms/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:gda="http://i-content.org/gda/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
 <rdf:Description
  "rdf:about="http://www.sw.it.aoyama.ac.jp/2005/hiroyuki/data.rdf"/>
  <dcterms:references df:resource=
   "http://www.sw.it.aoyama.ac.jp/2005/hiroyuki/hiroyuki.rdf">
 </rdf:Description>
 <rdf:Description
  rdf:about="http://www.sw.it.aoyama.ac.jp/2005/hiroyuki/
  hiroyuki.html">
  <dc:creator>阿部 裕行</dc:creator>
  <dc:date>2005-11-24T18:22:50+0900</dc:date>
  <dc:title>Martin 研究室 阿部 裕行 (Hiroyuki Abe)</dc:title>
  <dc:description rdf:parseType="Literal">
   <gda:gda>
    <gda:su>
     <gda:adp>
      <gda:adp>
       <gda:np bfm="構造" prn="コウゾウ" tf="1">構造</gda:np>
       <gda:ad bfm="と" prn="ト" sem="並立助詞">と</gda:ad>
      </gda:adp>
      <gda:ad>
       <gda:np bfm="自然" prn="シゼン" tf="1">自然</gda:np>
       <gda:n bfm="言語" prn="ゲンゴ" tf="1">言語</gda:n>
            :(省略)
       <gda:v bfm="できる" prn="デキル">できる</gda:v>
      </gda:v>
      <gda:ad>
       <gda:np bfm="環境" prn="カンキョウ" tf="1">環境</gda:np>
       <gda:ad bfm="を" prn="ヲ" sem="格助詞">を</gda:ad>
      </gda:ad>
     </gda:adp>
     <gda:v>
      <gda:np bfm="構築" prn="コウチク" tf="1">構築</gda:np>
      <gda:v bfm="する" prn="スル">する</gda:v>
     </gda:v>
    </gda:su>
   </gda:gda>
  </dc:description>
 </rdf:Description>
</rdf:RDF>
```

図 2: アノテーションデータの例（一部省略）

3.1 重要語抽出

各名詞節の属性 `tf` に単語出現頻度を表す TF 値を算出してある．その数値を用いて TF–IDF という指標で重要語を抽出する．あらかじめ算出してある数値の使用で素早く重要語を抽出できる．次節以降の処理でも重要語抽出処理を内部的に用いている．

3.2 要約

アノテーションに対する要約は，その概要だけ知りたい様な場合に有効である．そこで，以前我々が提案した重要文抽出と文内要約を併用した手法 [4] を用いて，文章を任意に設定した要約率で要約する．重要文抽出では，GDA で文を表す `<su>` 要素を一つの単位とし，表題 (`dc:title`) が一番重要な文であるという仮定を用いる．GDA タギングした表題内の各要素と各文要素内の各要素を比較し，各文要素の重要度，文要素間の関連度を求める．それらの値と要約率をもとに，重要でない文要素を削除する．文内要約では GDA の情報を利用して得られる文のテキスト構造から各要素に非重要度を求め，その値，要約率，さらに経験則を適用し要素を削除する．この 2 つの処理を同時または別々に行い要約の精度を上げる．

3.3 クラスタリング処理

アノテーションが増加すると，それらをグループ分けする処理が必要となる．そこで以前我々が提案した手法 [5] を用いて，アノテーション集合に対してクラスリング処理を行う．クラスタリング対象のアノテーション集合からアノテーション内に出現する名詞要素（`<n>`，`<np>` など）の単語を抽出し，各アノテーションをクラスタと見なした単語ベクトルを生成する．ベクトルの比較でクラスタ間の類似度を求め，最も類似度の高い 2 つのクラスタを一つのクラスタに統合する操作を繰り返し，クラスタ集合を生成する．統合されたクラスタ集合の中から最終的に作成するクラスタを選別し，各アノテーションのベクトルの和からクラスタラベルを決定する．

4 構造マイニング

図 1 のグラフ構造に対しては，構造マイニングを行うことができる．グラフ構造内から重要なアノテーションを判定できることは有用である．今回は，活性拡散という手法 [6] を用いて各アノテーションの重要度を得る．アノテーションに相当するノードに対して，全てのノードから活性拡散を行うとする．活性が拡散するとは，ある活性化したノードから，リンクで結ばれている隣のノードに活性を伝播させることを指す．この際，活性元のノードに入ってきた値を活性先のノードに加算する．3.1 節で解説した手法で算出したアノテーションの重要度が活性値となる．

5 まとめ

アノテーションの構造化に関する一方式及び，それに基づく自然言語処理と構造マイニングによる処理手法を提案した．これに基づいたアノテーション解析処理ライブラリを Java で構築した．このライブラリは近日中に http://www.sw.it.aoyama.ac.jp/2005/hiroyuki/ 上で公開する予定である．このライブラリを利用し，関連文生成システムを構築している [7].

参考文献

[1] Dave Beckett: RDF/XML Syntax Specification (Revised)，W3C Recommendation 10 February 2004. http://www.w3.org/TR/rdf-syntax-grammar/

[2] DCMI Usage Board: DCMI Metadata Terms，2005-06-13. http://dublincore.org/documents/dcmi-terms/

[3] 橋田浩一: Global Document Annotation (GDA)，草稿 第 0.74 版 (2005 年 10 月 17 日）. http://i-content.org/gda/tagman.html

[4] 伊藤一成，酒井康旭，斎藤博昭: メタデータ解析と自然言語処理を併用した要約動画の生成，情報処理学会研究報告，DBS-132，pp. 41-48，2004.

[5] 滝本湖，伊藤一成，斎藤博昭: 汎用アノテーションシステム（MAML System) を利用した Web 検索結果のグラフ表示，データベースワークショップ，DBWS2005.

[6] J. R. Anderson: A Spreading activation theory of memory，Journal of Verbal Learning and Verbal Behavior，pp. 261-295，1983.

[7] 望月英樹，阿部裕行，伊藤一成，Martin J. Dürst: Google API を用いた関連文生成の一手法，第 68 回情報処理学会全国大会，2006.