検索のための広告文書構造化

1V－4

井上香織　高橋克己
NTT ソフトウェア研究所

1．はじめに ―検索と文書の構造化―

現在Web上の文書検索の多くが全文検索だが，単語の文脈中の意味（属性）を考慮していないため，無駄な検索結果を多く返す．本研究は，Web上の雑多な広告文書の構造化を自動で行い，属性を指定した検索を可能とすることを目的とする．

ここで構造化とは，文書内の情報を属性とその値の集合に変換することをいう．従来の自動文書構造化の手法は次の二つに分けられる．一つは，特定の情報を表す言語表現をパターン化することで，その情報を抽出するものである[1][2][3]．特定の属性を決定する．もう一つは，形式が一定のパターンで繰り返されているリスト形式の文書について，その形式をパターン化し，不特定の属性を推定するものである[4]．特定の形式の文書を対象とする．しかし，Web上の雑多な広告文書を一律に構造化するためには，不特定の属性を不特定の形式の文書から決定することが不可欠である．

2.　広告文書の特徴

形式不定の文書から属性を決定するためには，文や単語の配置等の形式的な解析ではなく，意味的な解析をする必要がある．テキスト情報には，動詞等を含む普通の文章（自然文）と単語の羅列（単語列）があるが，前者と後者では単語の文脈上の意味の解析方法が異なる．自然文であれば動詞中心の格文法を用いた解析などが有効であるが，単語列には適用できない．よって自然文と単語列の割合を調べた．（表1）

2.1　形式調査

Yahoo Japan[7]のビジネスのカテゴリから30543の広告文書を集めた．各文書を1行毎に形態素解析し，動詞，格助詞，助動詞を含む行を自然文に分類，他を単語列に分類した．行とは，作者が改行を指定した部分までを指す．(
等)

表1　自然文の割合別文書数と全広告文書中の割合

自然文の割合	文書数	全文書中の割合（%）
3/4 以上	2377	7.7
2/4 以上 3/4 以下	6027	19.7
1/4 以上 2/4 以下	9096	29.7
1/4 以下	13043	42.7

2.2　結果

表1は，1文書内の自然文行の割合ごとに文書の合計数と全文書中の割合を示している．自然文より単語列のほうが多い文書が7割以上であった．この結果，単語列解析手法の検討の必要性が確認できた．

3．提案する構造化手法

2章では，単語列に適し，かつ意味情報を用いた構造化手法が必要であることを述べた．さらに不特定の属性に対応するため，次の条件をたてる．

・不特定の属性に対応できるだけの量の判別ルールがあること

以上の条件より，属性を識別できるような単語（Token）の辞書を用いた手法を提案する．Token辞書には図1のように，Tokenとその値となる単語が記述されている．辞書と文書内の自立語のマッチングによりTokenを抽出する．辞書に値とし

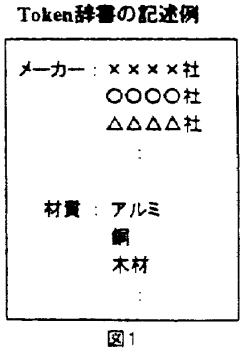

図1

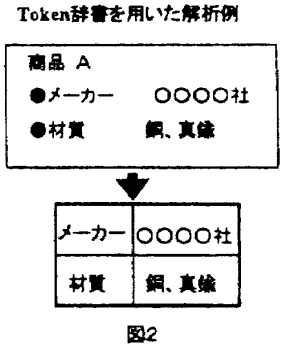

図2

Structuring of advertisement texts from the Web
Kaori Inoue　Katsumi Takahashi
NTT Software Laboratories

て記述されている単語がTokenと共起していれば，Token とその共起単語を，それぞれ属性と値に認定する．（図2）

Token辞書のTokenと値の単語には階層関係がある．よって，既存の辞書[7][8]の概念木を応用してToken辞書とすることも考えられる．しかし，単語間の関係は分野によって異なることが多い．また，膨大な固有名詞に対応する必要がある．従って，既存の辞書は，そのまま応用できないため，既存辞書の単語間の階層関係に対し，分野に特化した改良を行う．分野知識を抽出するために，HTMLタグの情報を利用する方式を検討する．

4. HTMLタグ調査

HTML タグのうち，属性とその値を表すようなタグは，表，リストおよび，フォントのスタイルタグである．HTML タグを利用した辞書作成の可能性を探るために，これらのタグの出現頻度を調べた．

対象

自然文が1/2以下の文書数	22139
開始タグを含む全行数	1285549
うち単語列行数	909110
うち自然文行数	376439

表2 単語列文書中のHTMLタグ割合

	タグ数	割合 %
td	368826	40.3
tr	144577	15.9
b	91504	10.0
table	64060	7.0
li	22634	2.4
th	20372	2.2
strong	9240	1.0
h3	7311	0.8

表3 自然文文書中のHTMLタグ割合

	タグ数	割合 %
td	97096	25.7
tr	49848	13.2
b	37930	10.0
table	25651	6.7
li	25201	6.6
dd	6161	1.6
ul	5514	1.4
h3	4247	1.1

表2，3より以下のことが言える．

- 単語列においては，表の要素が半分近くあり，表の解析が有効．(td，th，li，b，strong，hの合計より)
- 表，リスト，見出しタグ等の情報を用いて属性抽出ができれば，全単語列中の半分は解析可能．
- 自然文においても，3割程度はこの方法で解析可能．

5. HTMLタグを利用した推定

以上の調査で，属性を表すHTMLタグの頻度は十分得られることが分かったので，タグの情報を利用して単語の階層構造を抽出するツールを試作し，本手法の有効性を検証した．評価は，サンプルデータに対して，属性と値を表しているかを人手により判断し，適合率を求めた．

5.1 実験

表タグ，リストタグについて，対応する文字列を抽出し，属性と値の対応を示すように出力した．実際の出力例を示す．（図3）

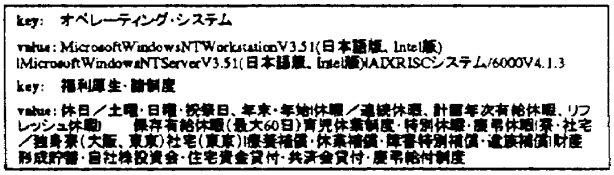
key: オペレーティング・システム
value: MicrosoftWindowsNTWorkstationV3.51(日本語版、Intel版)
MicrosoftWindowsNTServerV3.51(日本語版、Intel版)AIXRISCシステム/6000V4.1.3
key: 福利厚生・諸制度
value: 休日／土曜・日曜・祝祭日、年末・年始休暇／連続休暇、計画年次有給休暇、リフレッシュ休暇　慶弔有給休暇(最大60日)育児休業制度、特別休暇、慶弔休暇(家・社宅／独身寮(大阪、東京)社宅(東京))慶弔補償・休業補償・障害特別補償・遺族補償財産形成貯蓄・自社株投資会・住宅資金貸付・共済会貸付・慶弔給付制度

図3 出力例

現段階では，適合率は約3割であるが，ルールを改良することでより高い精度の階層構造抽出が見込める．また，フォントスタイルタグについても同様の実験を行う予定である．

6. まとめ

広告文書構造化のため，単語列解析の必要性を明らかにし，Token辞書を用いた手法を提案した．さらに，辞書の自動作成のため，HTMLタグの情報を利用することの有効性を確認した．

＜参考文献＞

[1] 佐藤他
"電子ニュースのダイジェスト自動生成"
情報処理学会論文誌 Oct 1995 Vol.136 No.10

[2] Jerry R. Hobbs et.al
"FASTUS: A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text"
http://www.ai.sri.com/~appelt/fastus.html 1996.

[3]J.Hammer et.al
"Extracting Semistructured Information from the Web"
Workshop on Management of Semistructured Data 1997.
http://www.research.att.com/~suciu/workshop-papers.html

[4] Naveen Ashish and Craig Knoblock
"Wrapper Generation for Semi-structured Internet Sources"
Workshop on Management of Semistructured Data 1997.

[5] 池原等 "日本語語彙大系" 岩波書店 1997.

[6] EDR 電子化辞書 http://www.iijnet.or.jp/edr/

[7] yahoo Japan http://www.yahoo.co.jp