ブートストラッピングによるスキーマ抽出のための キー抽出のチューニング

学籍番号：09C07110　　西田研究室　　野村慎太郎

1　はじめに

Web 上には大量の半構造化文書がある．しかし，これらはデータベースのように整形された形式で保存されていないため，計算機は正確に処理できない．この問題を解決するためには，半構造化文書からスキーマ（属性名の組）を抽出し，属性名に対応する属性値を抽出する必要がある．しかし，前者の処理について自動化を試みた研究はない．そこで本研究では，ブートストラッピングを用いて属性名を自動で獲得する手法を提案する．

ブートストラッピングとは，人手で属性値を少数与えておき，抽出値と抽出テンプレートを繰り返し学習することによって，少量の属性値から大量の属性値を抽出する手法である．

属性値の前後には，それに対応する属性名と，他の属性名が存在することが多いため，そのテキストの多様性は低い．そのため，前後のテキストが属性値の抽出テンプレートとして有効である．本研究では属性名を抽出しなければならないが，この手法をそのまま用いるには問題がある．属性名の前後には属性値が含まれることが多くなるため，その多様性は高くなるのである．多様性が高くなれば，各テンプレートが適合する確率が低くなり，Web 全体から属性名を抽出することは困難となる．

2　研究のアプローチ

ブートストラッピングでは，すでに抽出した値を次の抽出に利用できる．また，Web 全体で見れば各ページの記述形式の多様性は高いが，個々のページに着目すると，レコード中のデータ要素は同じ形式で繰り返し記述されていることが多い．したがって，2 つの属性名があれば，現在のページで有効となる属性名抽出用のテンプレートが学習できる．具体的には，すでに獲得している 2 つの属性名で Web 全体を検索し，適合するページにおいてそれぞれの単語の前後にあるテキストで完全一致する部分を現在のページに有効なテンプレートとする．

しかし，この方法で作成したテンプレートは，タグや記号だけから構成され，短いものであると考えられる．そのため，1 つのページ中でも様々な場所に適合し，目的とする属性名以外のテキストも抽出してしまう可能性がある．そこで，Web の文書構造を解析することにより，抽出範囲を 2 つの属性名が記述されたレコード部分に限定する．

また，属性名だけではオブジェクトのレコードが記述されたページだけでなく，用語解説のようなページも適合してしまう．そこで本研究では，2 つの属性名に加えて，キーとなる属性値（例えば，パソコンなら製品名．以下，キー）も検索することとする．また，決まったキーだけを用いていると，多様なスキーマを抽出することができない．そこで，キーもブートストラッピングにて学習する．

3　提案手法

抽出処理の流れを図 1 に示す．左側がキー抽出，右側が属性名抽出の流れを表す．

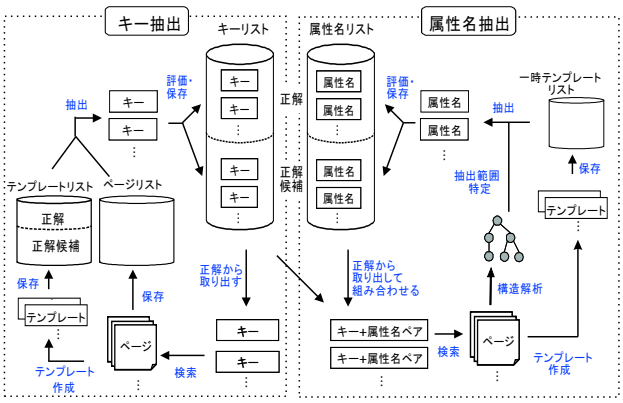

図 1: 処理の流れ

キー抽出処理については，前後の単語を抽出用テンプレートとし，獲得したページから抽出を行うものとする．また，誤った語が辞書に混入することを防ぐために，Yangarber らが用いた確信度により抽出値をスコア付けし，評価を行う．

4　キー抽出実験

本論文では属性名抽出の手掛かりとなるキー抽出に着目する．本実験の目的は，精度・抽出数が最も良くなるキー抽出用パラメータの組を調査することである．キー抽出用パラメータとは，テンプレートに用いる単語数，各サイクルで正解として保存するキーの更新数，テンプレートの更新数，の 3 種類を指す．以下表 1 に，(単語数，キー更新数，テンプレート更新数）の組の抽出数に対する精度を示す．

表 1: 各パラメータの抽出数に対する精度

	抽出数				
param.	20	60	100	200	260
(2,10,50)	0.650	0.667	0.620	0.605*	0.635*
(3,10,50)	0.850*	0.650	0.580	0.520	0.512
(2,2,50)	0.800	0.650	0.560	0.520	0.438
(2,5,50)	0.800	0.683*	0.670*	0.565	0.527
(2,20,50)	0.700	0.550	0.480	0.355	0.319
(2,10,40)	0.600	0.633	0.540	0.440	0.446
(2,10,100)	0.400	0.267	0.160	0.110	ー
(2,10,200)	0.400	0.233	0.140	0.0850	ー

*マークは最も良い精度を持つパラメータの組を表す

抽出数が 260 の時，精度は（2，10，50）の組が最も良いことがわかった．しかし，抽出数によって最大の精度を持つパラメータの組は異なる．

5　今後の予定

抽出したキーを用いて，属性名抽出実験に着手し，最良の属性名抽出用パラメータの組を決定する．