8ZA-6

自動情報収集による商品選択支援システムの提案

小島 岳史† 清光 英成†† 田中 克己‡

† 神戸大学工学部 †† 神戸大学大学院経済学研究科 ‡ 神戸大学大学院自然科学研究科

1 はじめに

ユーザーはWebから所望の情報を得るために，検索エンジンを用いて複数のページを閲覧し，それぞれのページの内容を比較して最も相応しいデータを選択する．現状のWebシステムでは，内容の判別と適合データの比較をユーザーが手動で行っている．そこで，本論文では自動的にユーザー所望の情報を提示する手法を提案する．

同属情報をもつWebページは，文章の構成やタグ構造が似ているという仮定に基づいて，Webページから同属情報が記載されている部分を文章や構造の類似性から判定し，自動的に収集する．また，データを分析して各ページ中のデータをOEMグラフで表し，近似的DataGideのアルゴリズムによって共通のスキーマを導出する．そして，導出したスキーマを用いてデータを比較し，ユーザーに提示する．

2 同属情報の自動収集と比較

同じ，あるいは類似の性質をもつデータを共通の側面から捉えることで得られる情報を同属情報と呼ぶ．例えば，何種類かのプリンタの性能を示す仕様のデータ集合である．同属情報を比較することで次のような情報を提示できる．

- 複数の同属情報で，ある属性集合について類似の値を持つ同属情報の集合を提示する．例えば，プリンタの性能という同属情報について，ある属性の数値が近いものを類似とすると定義した場合，600dpiの印字性能を持つプリンタの仕様を提示する．
- 類似な複数の同属情報から，特異な属性，または特異な属性を持つ同属情報を提示する．例えば，600dpiの印字性能を持つプリンタで，両面印刷が出来るプリンタの仕様を提示する．

同属情報の取得 同属情報をもっているページを探し出す．そして，同属情報をもつページは文章の構成やタグ構造が似ているという仮定に基づいて，使われている単語の同意語，類義語やタグ構造から同属情報を判別・取得する．

情報の整理 取得したデータは，多くの場合サイトごとに項目の分け方や，使われている用語が違っている．このような半構造データから共通のスキーマを導出する手法としてDataGuide[1]がある．

DataGuideとは半構造データをOEMグラフで表したときに導出されるスキーマである．あるパスを経由して到達できる全てのオブジェクトをターゲットセットと呼ぶ．ターゲットセットが完全に一致したパスで示されるオブジェクトがスキーマのインスタンスである．

本研究では用語や構造の違いを吸収できる，近似的DataGuide[2]を用いる．近似的DataGuideは，厳密には同じでないターゲットセットを同じターゲットセットとして扱うDataGuideである．

比較と提示 ユーザーとのインタラクションによって，データを比較する基準を決定する．決定した基準によってデータの比較を行い，類似の同属情報の集合を得る．

また，森らの差別化アルゴリズム[3]を用いてデータの特徴を導出し，特異な点，共通な点を得る．差別化アルゴリズムと，OEMグラフの全てのオブジェクトについて最も出現頻度の高いものを発見し，その下層から他のデータとは異なった部分を見つける．

更に，ユーザーの履歴を元にその嗜好や興味をデータを比較する基準の一つとする．

3 商品情報への応用例

本節ではパーソナルコンピュータの外部記憶装置に応用した例を示す．

商品仕様の取得 メーカーのサイトなどの商品仕様が記載されているページからデータを取得する．多くの場合商品仕様は表や箇条書きで示されているので，ページのソースの $<table>$, $<td>$, $<tr>$, $<ul>$, $<li>$ などのタグと，単語の類似性を用いて商品情報を取得する．表1，2に示された仕様書は図1のようなOEMグラフになる．このシステムで作られるOEMグラフでは，枝についたラベルは仕様書の項目名をあらわし，葉ノードには項目のデータとしての値が入る．

情報の整理 商品仕様のデータから得たOEMグラフから，スキーマを導出する．ただし，ページごとに

"Proposal of commodity selection support system by automatic information gathering"
†Takeshi Kojima, Faculty of Engineering, Kobe University. ††Hidenari Kiyomitsu, Graduate School of Economics, Kobe University. ‡Katsumi Tanaka, Graduate School of Science and Technology, Kove University

表 1: 仕様書 1

項目		項目の情報
外部記憶	ＦＤＤ	3.5 型（1.44MB ／ 720KB）× 1
	ＨＤＤ	60GB
	ＣＤ－ＲＷ	読み出し 32 倍速，書き込み 12 倍速
ＨＤＤコントローラ		Ultra ATA（66）

表 2: 仕様書 2

項目	項目の情報
フロッピー	3.5 型（1.44MB ／ 720KB）× 1
内蔵ＨＤ	80GB Ultra ATA（66）
ＣＤ－ＲＷ	読み出し 32 倍速，書き込み 8 倍速

使われている用語の違いを解決するために，オブジェクトだけでなくラベルの名前や，OEM グラフの葉ノードに格納された値の単位にも注目してターゲットセットの同一性を判定するように変更を加えた近似的 DataGuide を用いる．図 2 に，図 1 の OEM グラフから導出されたスキーマを示す．

履歴の取得 ユーザーの過去の購入履歴，閲覧履歴を蓄積しておき，そこからユーザーの嗜好や興味を推定して比較提示の手がかりのひとつとする．

情報の比較と提示 ユーザーとのインタラクションを行い，ユーザーの望む比較を行ってその結果得られた情報を提示する．

説明文との関連付け ユーザーに商品についてより詳しい情報を提供するために，商品の詳細情報を提示する機能を追加する．商品情報のサイトから，商品の説明文を取得して，説明文中に使われている単語の類似度からその説明が仕様のデータのどの属性に対する説明であるかを判定する．

4 まとめ

本稿では，複数の Web ページから自動的にデータの収集・分析・比較を行ってユーザー所望の情報を提示する手法を提案した．本手法は，複数の Web ページから特定の主題についてのデータを自動的に収集して共通のスキーマを導出し，比較して，情報を提示する．また，商品情報に応用した例を示した．

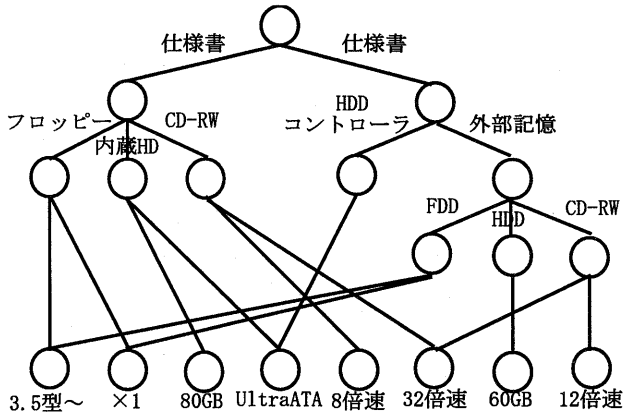

図 1: 仕様書 1，2 から得られる OEM グラフ

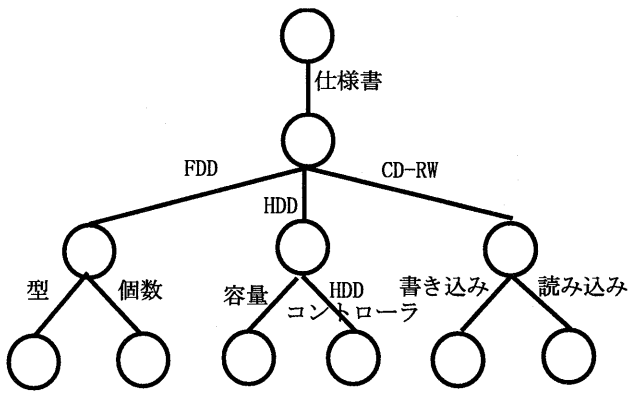

図 2: 図 1 の OEM グラフから導出されたスキーマ

ユーザー所望の情報を得るための支援をするだけでなく，類似な同属情報の中で，あるデータが他の情報と違う点を発見することから，ページ作者がページを改善する参考にできる．従って，本手法はユーザーとページ作者の双方にとって有用な手法であると言える．

参考文献

[1] Roy Goldman, Jennifer Widom: “DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases”, Proceedings lf the Twenty-Third International Conference on Very Large DataBases, pp.436-445, Athens, Greece, August 1997

[2] Roy Goldman, Jennifer Widom: “Approximate DataGuides”, Technical report, Stanford University, 1998

[3] Takehisa Mori: “Spatial Data Presentation and Incremental Query Formulation by Differentiation”, Master Thesis, Kobe University, February 2000