1M-4

レコード間の相関を用いたWeb情報統合方式の評価

于　美麗†　天笠俊之†　北川博之†

†筑波大学大学院システム情報工学研究科

〒305-8573 つくば市天王台 1-1-1

1　はじめに

近年，インターネットの発展により，様々な情報を入手することが容易になってきている．その中で，Web ブラウザを用いて，データベース情報を検索，ブラウズすることができる Web データベースが広く普及している．ショッピングサイト，論文共有サイト，旅行サイトなどが Web データベースの具体例としてあげられる．

このような有用なコンテンツを含む Web データベースに対して，複数の Web データベースを横断し，より複雑な問合せを行いたいという要求がある．例えば，テレビやカメラなど更新早い電子製品を買う時，コマーシャルの出演者は有用な参照情報になる．このような要求に対して，通常では商品情報を持つショッピングサイトと CM 出演情報を所持する製品のホームページをそれぞれ個別に問合せをして，人間が各 Web データベースの検索結果を見て判断する．これは大量のデータを効率的に処理することが難しいため，Web データベースを統合した上で問合せをすることが望ましい．

従来のデータベース統合ではレコードの持つ属性値の等価性を手掛かりに統合を行う．しかし，異種データベースにおいては，たとえ同じ属性であってもデータベースによって値の表現が違うことが頻繁に生じる．また，同一の属性により統合することだけではなく，ある種の関連性に基づいた関連付けを行いたいことも多い．上の例においては，「商品」と「人物」には CM 出演という関連性がある．本研究の目的はこのような関連性により異種 Web データベースの統合を行うことである．

一方，近年，ブログやツイッターなどソーシャルメディアが新たな情報源として注目されている．これらは情報の更新も速やかに行われるので，その情報を利用して明示的に示されていない物事間の関連性を見つけ出すことが可能である[1] [2]．このため我々は[5]において既存の検索エンジンを用いることで Web 上に存在している大量の情報からレコード間のアソシエーションを検出し，更に，検出したレコード間のアソシエーションを利用して Web データベースを統合するシステムを提案した．本論文では．提案したシステムの評価結果について述べる．

2　提案システム

本節では，我々が提案する Web 上の情報から検出したレコード間の相関を利用して，Web データベースを統合する手法を述べる[5]．システム構成を図 2.1 に示す．

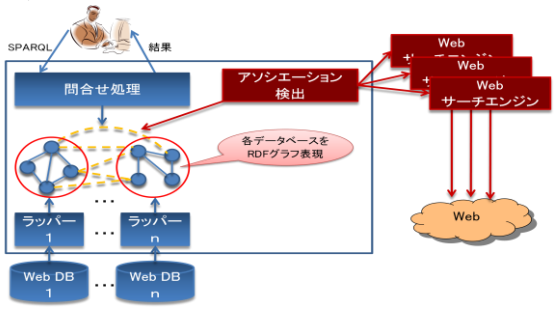

図 2.1　システム構成図

提案システムは，1）ラッパー，2）レコード間の相関検出，3）問合せ処理に分けられる．

2.1　ラッパー

本研究では，ラッパーを用いて各 Web データベースを RDF[3]データとして扱う．これは，RDF が任意の述語を定義可能であり，多種多様な Web データベースを統一的に表現可能なためである．Web データベース毎にラッパーを作成する必要がある．ラッパー作成の手続きは下記のようになっている．

(1) Web データベースから抽出の対象とするレコードやレコードの属性を選定する．
(2) 選定したレコード及びレコードの属性により，RDF スキーマを定義する．
(3) レコードのキー属性を決める．
(4) リソースに付与する URI の自動生成規則を制定する．

作成したラッパーは RDF スキーマに従って Web データベースを RDF グラフに変換する．

この規則による，図 2.2 で示した Web データベースが，図 2.3 のように変換される．

図 2.2　Web データベースの例

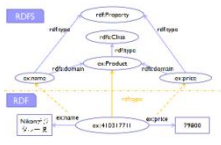

図 2.3　変換された RDF の例

2.2　レコード間の相関検出

ラッパーから得られた RDF データは，データベース毎に閉じているので，Web 検索エンジンを利用して，Web データベースの各レコード間の相関を検出する．具体的な手順は以下の通りである．

レコード A と B との間の関連度を推測したい．ただし，レコード A はリテラルのキー属性 P_A，レコード B はリテラルのキー属性 P_B を含むとする．

An evaluation of the scheme for Web database integration using associations among records

Meili YU† (yumeili@kde.cs.tsukuba.ac.jp),
Toshiyuki AMAGASA† (amagasa@ cs.tsukuba.ac.jp),
Hiroyuki KITAGAWA† (kitagawa@cs.tsukuba.ac.jp)
†Graduate School of Systems and Information Engineering, University of Tsukuba

(1) 検索エンジンに対して，属性 P_A と P_B をそれぞれ単一のキーワードとして与えたときのヒット件数（$|P_A|$，$|P_B|$）を得る．
(2) P_A と B と同時に検索キーワードとしたときの検索エンジンのヒット件数（$|P_A \cap P_B|$）を得る．
(3) 閾値つき Simpson 係数を用いて，レコード A と B との関連度（R(A, B)）を計算する．

$$R(A,B)=\begin{cases}\dfrac{|P_A \cap P_B|}{\min(|P_A|,|P|)}, & \text{if } |P_A| > k \text{ and } |P| > k \\ 0, & \text{otherwise}\end{cases}$$

2.3 レコード間の相関の RDF 表現

算出されたレコード間の相関は一種のリソースと考えられるので，RDF トリプルで表現する．そのため，レコード間の相関を表記する RDF スキーマを定義する(図 2.4)．

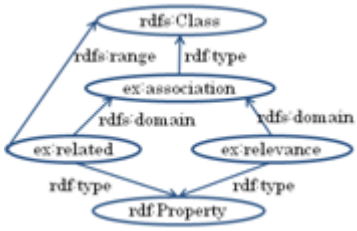

図 3.6 レコード間の相関の RDF スキーマ

クラスアソシエーション（ex:association）はレコード間の相関を表すリソースタイプであり，二つのプロパティを持つ．レコード間の関連度を表すプロパティ（ex:relevance）と関連するレコードを表すプロパティ(ex:related)である．

2.4 問合せ

レコード間の相関を追加することで統合した Web データベースを検索するため，RDF 問合せ言語 SPARQL[4]を利用する． 問合せ処理でユーザの検索要求により SPARQL クエリを生成し，統合した RDF データから検索して，結果をユーザに返す．

3 評価実験

本節では提案システムの有効性を実験により検証する．実験環境は次のようになる．

CPU	Intel Core2 Duo 1.40GHz
メモリ	2GM
OS	Windows Vista Business SP2
ランタイム	JRE 1.6.0_23

表 3.1 実験環境

実験データは Wikipedia から抽出した 45 個の電機メーカー及び製品情報と 45 人の人物情報である．実験手法としてはメーカー情報と人物情報を統合するために必要な処理時間を計測するとともに，検出したレコード間の相関の正解率を計測した．さらに，統合したデータに対して，図 3.3 で示したクエリを発行し，検索時間を計測した．検索エンジンの Yahoo! Search BOSS を利用した．また，閾値つき Simpson 係数の閾値は k=30 とした．

検索結果を図 3.4 に示す．相関検出を含めた相処理時間は 1403.425 秒であった．これは，主に検索エンジンへの問合せに費やされていた．

次に相関検出の正解率は図 3.2 に示す．横軸はレコード間の相関を判定する閾値であり，縦軸は閾値の変化に伴い抽出した相関の数と相関の正解率である．閾値の増加と共に検出した相関の件数が下がる傾向である．閾値が 0.05 の時に正解率が最も高いことが図 3.2 から分かる．

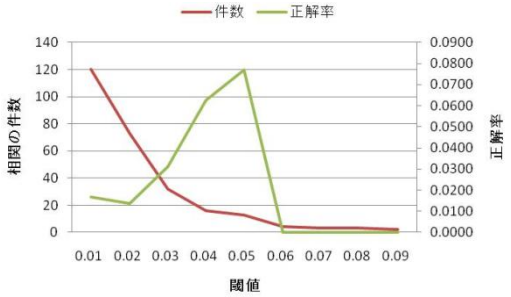

図 3.2 相関抽出の件数と正解率

```
1. SELECT ?product ?companyName ?personName
2. WHERE{
3.    ?company exC:product ?product.
4.    ?company exC:name ?companyName.
5.    ?asso exA:related2 ?company.
6.    ?asso exA:related1 ?person.
7.    ?person exP:name ?personName.
8. FILTER regex(?product, "炊飯器")
9. }
```

図 3.3 検索クエリ

商品名	メーカー	人物
炊飯器	タイガー魔法瓶株式会社	中山エミリ
炊飯器	タイガー魔法瓶株式会社	水川あさみ
・・・	・・・	・・・

図 3.4 検索結果

検索時間は 5 回平均で 0.4325 秒であった．システムとしては十分な速度であると考える．

4 まとめと今後の課題

本研究では，Web 検索エンジンを利用してレコード間の相関を発見し，それによって複数の Web データベースを統合するシステムを提案した．実験により，提案手法の有効性を示した．

今後，統合処理の効率化を計りより大規模なデータへの適用を試みる予定である．また精度の向上と，レコード間の相関を単なる相関の存在だけではなく，Web の検索結果から導出することなども今後の課題である．

謝辞

本研究の一部は科学研究費補助金特定領域研究（＃21013004）による．

参考文献

[1] 松尾豊，友部博教，橋田浩一，中島秀之，石塚満．Web 上の情報からの人間関係ネットワークの抽出．人工知能学会論文誌 20 巻 1 号， 2005
[2] Makoto P. Kato, Hiroaki Ohshima, Satoshi Oyama, Katsumi Tanaka. Query by analogical example: relational search using web search engine indices. CIKM, 2009
[3] RDF. http://www.w3.org/TR/rdf-primer/
[4] SPARQL. http://www.w3.org/TR/rdf-sparql-query/
[5] 于美麗，天笠俊之，北川博之． “アソシエーション抽出を用いた Web 情報の統合方式”，情報処理学会第 72 回全国大会講演論文集(1)，pp.639-640