B-035

博物館情報におけるメタデータスキーマ間の類似性に関する一考察

A Study On Similarity Between Metadata Schema Based on Museum

秋元 良仁 †　　亀山 渉 †
Ryoji Akimoto　Wataru Kameyama

1 はじめに

情報技術の普及に伴い，人類によって創出される情報量は爆発的に増加している．そのため，膨大な情報から利用者にとって有益な情報を整理提供する技術の確立が急務となっている．この技術の確立には 2 つのアプローチが考えられる．一方はメタデータを用いた情報整理技術 [1]，もう一方は統計解析手法を用いた情報整理技術である．前者はメタデータを用いてデータを効率的に管理し，利用者はそれを介して目的の情報源にアクセスする．後者は語の出現頻度やリンクの関係情報等を用いてデータを解析し，利用者は解析結果を介して目的の情報源にアクセスする．

筆者らはこれまで前者の立場から，メタデータを用いた情報の整理技術について論じてきた [2]．メタデータは様々な用途に応じてスキーマが定義されるため，スキーマ間の関係性が不明瞭であり，横断的な情報利用が困難な状況にある．そこで，複数の異なるスキーマ間の関係性を記述できる言語 Fuzzy Schema を提案している．現在, 博物館分野に存在する複数のスキーマに対し，手作業による Fuzzy Schema 記述で関係性が記述できることを確認しているが，今後，様々なスキーマに Fuzzy Schema を適用させるには，自動的に Fuzzy Schema を生成させる必要がある．

本稿では，メタデータ構築に関するいくつかの技術，統計解析に関するいくつかの技術を概観し，比較整理を行う．その上で Fuzzy Schema を自動生成するためのアプローチを考察する．

2 メタデータ構築技術

2.1 Heavyweight メタデータ

Semantic Web は Web 上のリソースに機械可読な意味付けをすることで人間に代わりソフトウェアが自動処理する技術の総称を言う．リソースの持つ属性と値の関係を記述する RDF, その語彙定義に当たる RDFS, RDF で定義される概念間の関係性と語彙を定義する OWL，その推論ルールを規定する SWRL 等，高機能言語が標準化されている．これらの言語を用いて知識体系を厳密に定義することでアプリケーションはスキーマで定義されたメタデータ間の自動処理が可能となる．しかし，知識体系の構築には専門家の手間と時間を要する．更に，構築した知識は日々出現する新しい知見や誤り等によりメンテナンスされて行くことが前提となる．

2.2 Lightweight メタデータ

高機能言語を Heavyweight(重量級) なメタデータと呼ぶならば，blog で広く用いられる RSS，XHTML を独自拡張した形式でリソースにメタデータを付与する Microformats や XHTML2.0 仕様に準拠した形式で RDF を付与する RDFa のようなメタデータの実装法は Lightweight(軽量級) なメタデータと呼ぶことができる．これらのメタデータは特定の問題解決やアプリケーション開発が容易であるというメリットを持つ反面，メタデータ間の語彙の差による混乱や概念の衝突といったデメリットを持つ．

2.3 人手によるメタデータ

リソースへのアクセスを補助する手段として，Folksonomy と呼ばれるメタデータの分類法がある．Folksonomy は folks(人々の) と taxonomy(分類学) を組み合わせた造語で，あるサービスを利用するユーザがそのサービスで使用されるコンテンツに対して個々に tag と呼ばれるキーワードを付与することでコンテンツを分類する手法を言う．キーワード自体は何ら体系立てて付与されないが，ユーザ数の増加に伴いコンテンツに付与されたキーワードがコンテンツに関する記述 (メタデータ) としてあいまいに深みと広がりを持って行く．思いがけない発見ができる反面，ユーザの主観に左右される場合が多く，ノイズ情報も多くなる．

3 統計解析技術

3.1 文献計量学的手法

文献計量学は文献の語の数量的特徴を分析する学問であり，古くから論文分析等で活用されてきた．近年では，応用例の 1 つとして Web ページの重要度を測る手法である Google 社の PageRank がある (1).

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \qquad (1)$$

N は対象 Web サイト数，$PR(p_j)$ はページ p_i にリンクしているページ p_j の PageRank，$L(p_j)$ はリンク元のページ P_j に含まれる他サイトへのリンク数を表す．PageRank は Web ページの意味内容を取り扱わず，ページの被リンク量によって当該ページの重要度を算出する．そのため，情報整理に人間のあいまいな知識表現を必要としない．しかし，リンクの意図まで汲み取ることはできず，重要度の信頼性を確保することは難しい．

3.2 マイニング技術

テキストマイニングは文章のような半構造化／非構造化データから語の出現頻度や相関関係を計算することで知識発見を行うデータマイニング手法の 1 つである．近年では Web 上に存在する様々なコンテンツ (フォーマット，リンク，ログ，クッキー等) を用いて知識発見を行う Web マイニング技術，リンクのグラフ構造や Web コミュニティでの関連性発見，アクセスパターンや動作履歴解析技術等，情報源の確率モデリング研究もなされて

† 早稲田大学大学院国際情報通信研究科

表 1　従来技術の比較

技術	知識の利用・発見			知識の構築・維持	
	知識の発見	知識の関係性利用	知識の信頼性	知識構築負荷	知識メンテナンス負荷
Heavyweight メタデータ	○	○	△	×	×
Lightweight メタデータ	○	○	△	△	△
人手によるメタデータ	△	△	△	○	○
統計解析	△	×	×	○	○

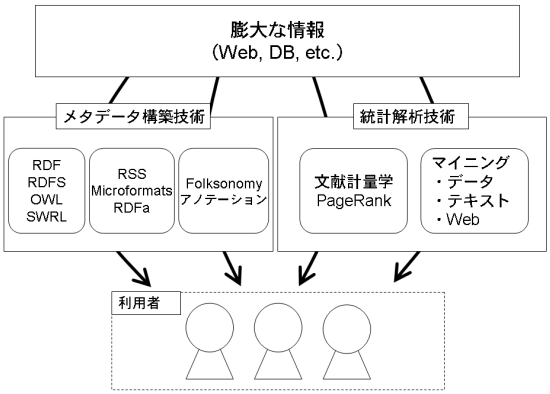

図 1　膨大な情報の利用技術

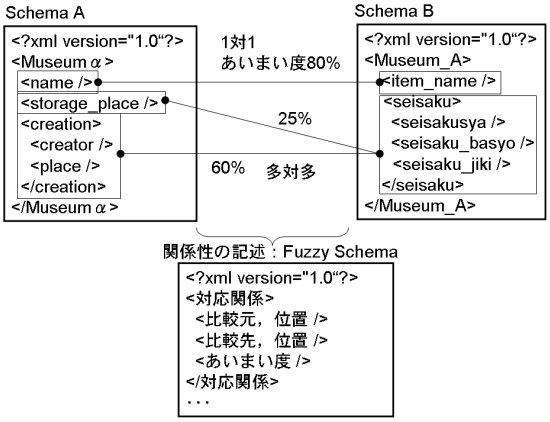

図 2　Fuzzy Schema

いる．これらの研究は，情報源を定量的に解析できるため，データの規則性や様相を見つけることはできるが，情報源の利用には人間の解釈を要する．

4　従来技術の比較

これまで概観してきた技術を図 1 にまとめ，その利点・欠点を表 1 にまとめる．

表 1 は，各技術を知識の利用・発見，構築・維持の観点から整理している．

メタデータを用いる場合，知識間の関係性記述が可能であり，知識発見も行われやすい．また，メタデータで効率的に情報を整理するため，目的の情報源にたどり着きやすく，ある程度の信頼性も確保される．更に，人手によるメタデータ付与に見られるような思いがけない知識発見も見込まれる．その反面，知識を構築したりメンテナンスする必要があり，非常に負荷が高い．

統計解析手法を用いる場合，人間が構築した知識体系を利用しないため，その構築負荷やメンテナンス負荷は低い．その反面，知識発見には解析結果を人間が解釈する必要があるため，直観的な知識発見は難しい．また，解析結果にノイズが含まれることも考えられるので，信頼性は低いものとなる．

5　Fuzzy Schema 自動生成のアプローチ

Fuzzy Schema はメタデータを用いた情報整理技術の 1 つであり，複数のメタデータスキーマの項目間の対応関係が記述できるマッピング・パターンとマッピング・パターンで対応付けられる項目間の類似度合を示すあいまい度から構成される XML 形式の言語である (図 2)．筆者らはこれまでに手作業により Fuzzy Schema で対応関係の記述を確認をしているが，今後，多様なスキーマに Fuzzy Schema を適用するには Fuzzy Schema の自動生成が必要であると考えている．

表 1 はメタデータ構築技術と統計解析技術はトレードオフの関係であることを示している．そこで，双方の技術を相補的に組み合わせることで Fuzzy Schema の自動生成のアプローチを考える．例えば，2 つのスキーマでそれぞれ定義付けられたメタデータの要素名や属性名をラベルとして，ラベル間の共起情報を用いた統計解析が考えられる．この際，補助的にそれぞれのスキーマで定義付けられたメタデータで構成されるインスタンスを用いることで複数項目間の関係性の把握と類似性を計算することができる．

6　まとめ

本稿では，スキーマ間の関係性が記述できる言語 Fuzzy Schema の自動生成について考察するために，メタデータ構築技術，統計解析技術の整理比較を行った．今後は具体的な Fuzzy Schema 自動生成方式の実証に取り組んで行く予定である．

参考文献

[1] Dempsy *et al.*: “Metadata: A Current View of Practice and Issues”, J. of Documentation, Vol. 54, No. 2, pp. 145-172(Mar. 1998).

[2] 秋元良仁, 亀山渉: “博物館情報を用いたメタデータスキーマ統合機構の実装と評価”, 情処研報, Vol. 2007(2007-07).