テキストの類似性を利用したXML文書統合手法の検討

秋永　智†　木村　昌臣‡

芝浦工業大学大学院工学研究科電気電子情報工学専攻†　芝浦工業大学工学部情報工学科‡

1. はじめに

近年，XML形式のデータが様々な分野で利用されている．RSSやAjaxなどはXMLの技術を利用して作られており，またWeb上ではデータとしてのXMLが多数公開されている．一方でXML形式のデータが増えたことにより，類似した情報であるにも関わらず構造の違いや，記述されている要素の種類が多様化したことで統一的に扱えないという問題が生じている．そのため，異なるXMLをより有用な情報とし効率的に活用するため，類似したXML同士の統合が重要である．XMLの統合手法として木構造を用いた統合方法[1,2]があるが，文献[1]では実行時間の問題が，文献[2]では精度面の評価が行われていないという問題がある．また，多数の要素から同じ情報を有する要素を見つけ，XMLを統合する手段は確立されておらず，実際に統合を行うのは困難である．

そこで本研究ではテキストの類似性に着目して要素群を分類し，同じ内容を有する要素を同定する手法及びそれを用いてXMLの統合を支援するシステムの開発を行う．

2. 提案手法

本研究では同じ情報を有する要素を同定するため要素中に記述されているテキストを利用する．文字列の傾向から類似している要素同士をまとめあげ，さらにその中から同じ情報を有する要素を同定する．本研究では，構造及び記述されている要素の異なるXMLとしてWebAPIから得られるXMLを利用する．比較する2つのXMLをX_a，X_bとし，XMLが持つ各要素をE_a,E_bと表す．

2.1 要素の取得とテキストのベクトル化

日本語のテキストは通常ひらがな，カタカナ，漢字などの文字種で構成されており，内容によって使用される文字種はある程度決まってくる．よって，同じ情報を有する要素のテキストであればその傾向も類似すると考えられる．文字列には表記揺れなども存在するため文字を「漢字」「英字」などの文字種に分類し，テキスト中にある各文字種の出現回数をベクトル値として扱う．各要素 $E \in X_a$, X_bが持つテキストのベクトルVを

$$V_E = (q_1, q_2, q_3, q_4, q_5, q_6)$$

とし，各q_iを順にひらがな，カタカナ，漢字，数字，英字，記号の出現回数として定義し，各要素中のテキストをベクトルに変換する．各要素は繰り返し出現するため，同名の要素中に記述されているもの全てを取得し，変換の対象とする．

2.2 類似要素の分類

まず，X_aの各要素に対してクラスタリングを適用する．クラスタリングは任意の集合を一定の基準の下で部分集合（クラスタ）に切り分ける手法であり，本研究ではその一種であるk-means法を適用することで要素をk個のクラスタに分類する．入力パラメータには2.1で作成したベクトルを用いて文字種の傾向が似ている要素群に分類する．

次に上記のクラスタを用いることでE_bの分類を行う．各要素E_bに対応するV_{Eb}に似たベクトルV_{Ea}が属するクラスタ中に同じ情報を有する要素E_aが存在すると考えられる．そこでX_aのクラスタリングで得られたクラスタC_kとE_bの類似度を求め，E_bと同じ情報を有する要素E_aが属する可能性の高いクラスタへE_bを分類する．C_kに属する要素E_aがn個あったとき，

コサイン尺度$\cos(E_b, E_a) = \frac{V_{Eb} \cdot V_{Ea}}{\|V_{Eb}\|\|V_{Ea}\|}$を用いて$E_b$と$C_k$の類似度を

$$S(E_b, C_k) = \frac{\sum_{E_a \in C_k} \cos(E_b, E_a)}{n} \cdots (1)$$

とし，各E_bに対して類似度を算出する．各E_bに対して最も類似度の高いクラスタへと分類を行う．

2.3 同じ情報を有する要素の同定

同じ情報を有する要素であれば記述されているテキストも類似していると考えられる．しかし要素中のテキストはXML作成者が自由に記述しているために同じ情報を記述していても表記の仕方が異なり，文字列単位の比較のみでは全ての要素において正しい同定を自動的に行うのは困難であると予想される．そこで自動同定を補うため手動同定システムを実装し，2種類で要素の同定を行う．

要素の自動同定にはN-gramを用いる．N-gramとはテキストをN文字単位で分解し，文字列の出現頻度を求める方法で，繰り返し出現する共通文字列の抽出に適している．これにより頻出文字列を求め，文字列をベクトルの変数，出現回数をベクトル値とみなし，2.2と同様にコサイン尺度を用いて要素間の類似度を求める．最も類似度の高い要素の組み合わせが同じ情報を有する要素であると判断する．さらに，上記で得られた結果に加え，ユーザーはGUIによりドラッグ&ドロップ等の操作を用いて手動で要素の関連づけを行う仕組みを実装した．これにより同じ情報を有する要素同士の確認や不要な要素の除外を行う．

3. システム概要

本研究で作成したプロトタイプシステムについて記述する．上記で同定した同じ情報を有する要素群を用いてXMLの再編成を行い，XMLの統合を図る．図1にシステムのGUIを示す．

本システムの流れは以下の通りである．まずファイルを読み込み，クラスタ数kを決定してクラスタリングを行う．通常，クラスタ数kは手動で決めなければならないが，x-means[4]を用いることでkを自動で決定することが可能である．2.3の自動同定で出力した結果と手動同定を用いて要素の関連づけと修正を行い，XMLの統合を行う．XMLの統合には，同じ情報を有する要素の関連づけからXMLのタグと

XML document integration method based on similarity of text in elements

† Satoshi Akinaga, ‡Masaomi Kimura

†Graduate School of Shibaura Institute of Technology

‡Shibaura Institute of Technology

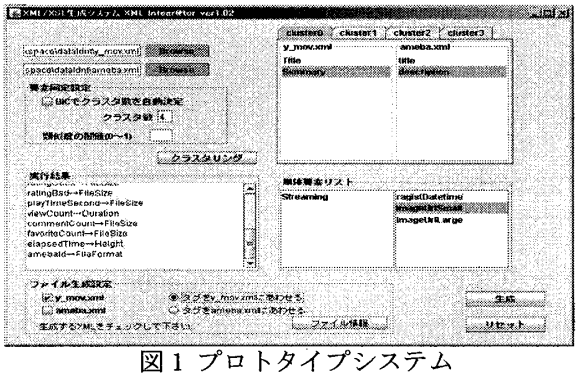

図1 プロトタイプシステム

構造を統一する XSL を自動生成する研究[3]を利用し XSL 及び変換後の XML ファイルを生成する．XSL を用いることで一度行った XML の組み合わせであれば同じ情報を有する要素の同定作業を省略することが可能となる．

4. 評価検証及び考察

4.1 実験概要

本手法の精度及び実行時間の評価を行う．実験は X_aにぐるなび Web サービス[5]（要素数 40），X_bにホットペッパー WEB サービス[6](要素数 68)を用い，実行環境は CPU：Pentium Dual-Core 1.8GHz，メモリ:1.5GB，OS：WindowsXP SP3 とする．精度は同じ情報を有する要素の正しい組み合わせを予め定義したものを正答とし，それと出力結果を照らし合わせ，同じ情報を有する正しい要素の組が同一クラスタに分類されていればクラスタ間同定の正答とし，システムが同じ情報を有すると判定した組と正答の組が一致していれば要素間同定の正答とする．また，実行時間については GUI の操作時間を除いた内部処理時間を，要素を同定する前半とファイルを生成する後半に分割して計測する．前半がファイルの解析とデータ取得，クラスタリングとクラスタ・要素同定時間の合計，後半が XSL ファイルと XML ファイル生成時間の合計に相当する．

4.2 結果・考察

クラスタ間同定の結果を表 1 に，実行時間の結果を図 2 に示す．

表1 クラスタ間同定の精度

クラスタ	精度	クラスタ	精度
1	0.85	5	—
2	1.00	6	1.00
3	1.00	7	0.80
4	0.66	8	0.85

上記の組み合わせによる実験ではクラスタ数は 8，クラスタごとの精度は表の通りになった．クラスタ 5 に分類された各要素は対となる XML に同じ情報が存在せず，クラスタ間同定に必要な前提を満たしていないため評価の対象外とした．クラスタ間同定全体の精度は 0.88，再現率は 0.9，F 値は 0.89 であった．上手く分類されなかった原因の 1 つとして表記の違いが挙げられる．例えばフリガナを意味する要素が存在するが，片方は「かな」表記，もう片方が「カナ」表記で記述されていた．また，長いテキストには記述の仕方に幅があったことが原因とみられる．要素間の同定は，クラスタ 2,3,7 においてほぼ 100%の同定精度が得られた．これらのクラスタは日本語で記述されていた要素群で，名詞などを上手く抽出できたことが要因であると考えられる．逆に 1,4,5,6 は正しい同定が行われなかった．これらは数字のみで構成される要素や，URL の要素であったがこれらのテキストを N-gram で抽出した文字列だけでは意味を持たず，共通した文字列が取得できなかったことが原因であると考えられる．

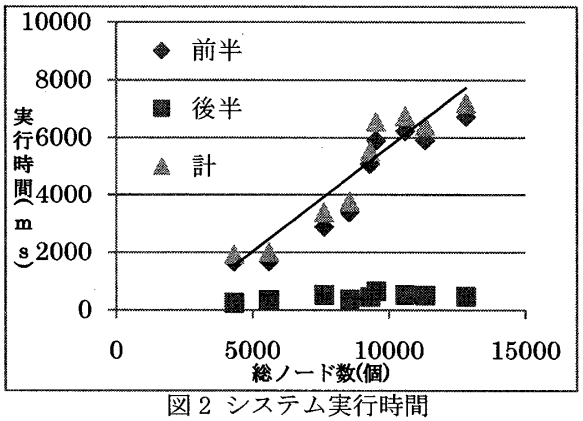

図2 システム実行時間

図 2 のグラフ中の直線は合計実行時間の散布図から最小二乗法を用いて得られた近似直線である．実行時間に関しては総ノード数が増加するにつれ実行時間も増加している．原因は N-gram 部分の文字列マッチングで時間がかかっていると考えられる．一方でファイルの生成時間はノード数にほとんど依存しないと言える．

5. まとめ

本研究では記述されているテキストの類似性に着目し，同じ情報を有する要素の同定手法を検討すると共に効率的な XML の統合手法について検討を行った．これにより適当な時間で実行が可能で，XML の統合が可能であることを示した．

今後の課題として実行時間の短縮や，元の XML の木構造に考慮した XML の統合が可能となるように検討を行う．

参考文献

[1] Lianzi etc:An approach for xml simi-larity join using tree serialization. In 13th Int'l Conf. on DSFAA (2008)
[2] 横田 他:複数の XML 文書の類似度検出方法および類似正検出システム，ならびに複数の XML 文書の統合手法，特許コード:P06P004283(2008)
[3] 秋永，木村:Web2.0 におけるマッシュアップ標準化手法の検討，情報処理学会第 70 回全国大会 5R-1(2008)
[4]石岡:クラスター数を自動決定する k-means アルゴリズムの拡張について，応用統計学 Vol.29, No.3, 141-149 (2000)
[5] ぐるなび Web サービス
http://api.gnavi.co.jp/api/service.htm
[6] ホットペッパーWEB サービス
http://webservice.recruit.co.jp/hotpepper/