

XML データにおける異種特徴量集約手法の検討

磯部 亮介[†] 福田 直樹[†] 石川 博[†]

[†] 静岡大学情報学部情報科学科 〒432-8011 静岡県浜松市中区城北 3-5-1

E-mail: cs04501@s.inf.shizuoka.ac.jp, {fukuta, ishikawa}@inf.shizuoka.ac.jp

あらまし Web 検索結果を、単一のページではなく、関連するページの集合（ページセット）として結果を返すことで、閲覧性の向上が期待できる。そのためには、相互に類似した半構造データを効率よく見つけ出す手法が必要となる。一方、Web ページの記述言語として、XML に基づく規格である XHTML が普及するなど、XML データが検索対象となる場面が増加してきた。従来の XML データの類似性の判定手法には、構造とテキストが主に用いられてきた。XML データには、画像や音声などテキスト以外のメディアが含まれる場合があり、類似性判定精度の向上にはそれらの異種メディアの類似度を考慮することが有益であると考えられる。本研究では、XML 内木構造と、テキストや画像などの要素値という 2 つの類似度に着目し、それらを集約することにより、XML データの類似性判定の精度を向上する手法について検討する。

キーワード XML, 類似度, 異種特徴量, 集約

An Examination of A Method Aggregating Different Characteristic Information on XML Data

Ryosuke ISOBE[†], Naoki FUKUTA[†], and Hiroshi ISHIKAWA[†]

[†] Department of Computer Science, Faculty of Informatics, Shizuoka University 3-5-1 Johoku, Naka-ku, Hamamatsu, Shizuoka, 432-8011 Japan

E-mail: cs04501@s.inf.shizuoka.ac.jp, {fukuta, ishikawa}@inf.shizuoka.ac.jp

Abstract On web-based systems, it is useful to respond results as sets of related pages that are called page sets. There is a need for a better algorithm to find similar semi-structured data from heterogeneous large data. On the other hand, a need for searching XML documents is increasing. An example is a need for searching XHTML that is a standard web page description language based on XML. In recent researches, structure-level and text-level measurements have been used to calculate similarity among XML data. However it is useful to consider similarities of different media for improving search precisions since there are some cases that XML data contains images, sounds, and other media data besides their texts. In this paper, we propose a method that improves precision of calculating similarities among XML data by using two similarities in tree structures and element values and aggregate them effectively.

Keywords similarity, XML, different characteristic information, aggregation

1. はじめに

近年、電子商取引や科学データなどのデータ記述、保存、交換の場において、XML により記述される半構造データの利用が拡大している。Web ページの記述においては、これまでの HTML によるものから、XML や XHTML によるものへと移行しつつあり、ページ全体に対するメタデータ記述を可能とする技術の研究が進められている。Microformats はその代表といえる [1]。半構造データの増加により、それらのデータの中から、必要なデータを適切に抽出する手法、高精度に検索結果を返す手法が要求されている。今後の Web 検索においては、ページ単位で検索に合致するページを検索結果として返すだけでは十分でなく、関連するページの集合を返すことが、利用

者にとって有用であると考えられる。

関連するページの集合を適切に求めるために、半構造データ間の関連性や類似性の効率的な発見手法が必要となる。一般的な半構造データの関連性の尺度としては、構造の同一性が用いられることが多い。また、Web ページ間の関連性の発見に最もよく利用されるのは、ページ間に張られたリンクである。ところが、Web ページには、図 1 に示すとおり、テキストや画像、音声などの異種メディアのデータが含まれる可能性が高く、それらの同一性もページの関連性の発見に利用できると考えられる。

そこで、本研究では、要素による木構造と、テキストや画像などの要素値という 2 つの特徴量に着目し、それ

```

<data>
  <text>宮崎県の知事は東国原氏です</text>
  <other>
    <image>Chiji.jpg</image>
    <movie>Chiji.mpg</movie>
    <sound>Chiji.mp3</sound>
  </other>
</data>

```

図1 複数種のメディアを含むXML文書

らを集約することにより、半構造データの比較・検索の精度を向上させる手法を提案する。

2. 関連研究

要素による木構造と要素値の両方を考慮した類似度の算出手法として、Ma と Chbeir は、同一のスキーマを持つXML文書について、要素値の類似度として語の意味的關係を利用して部分木の類似度を求めることにより、全体としての類似度を算出する手法を提案している[2]。また、文らは、XMLデータが持つ木構造処理の複雑さに対処するために、次のような手法を用いてXML文書の類似結合を提案している[3]。まず、要素を後置順で巡回して直列化する。その際、構造情報の保持のために、補助記号を導入している。その後、要素値を持つ要素の個数と部分木が持つ深さを閾値としてXMLデータを分割する。そして、部分木同士の類似度を計算する。類似度の計算には構造情報とテキスト情報が用いられ、テキスト情報についてはBloomフィルタ[4]を用いて各部分木の情報をビット列に変換することで類似度を算出し、この値が閾値以上のものについては、直列化された部分木同士の編集距離により構造情報についての類似度を算出する。集約という手法ではなく、段階を追って類似度を算出する点が、本研究で提案する手法と異なる点である。

異種メディア検索や異なる検索エンジンの結果を考慮する手法として、鈴木らは、Shannonの情報量の概念を利用したスコアの正規化、及び、各種評価関数によるスコアの統合により、集約された類似度を算出している[5][6]。鈴木らは、メタ検索エンジンの精度向上における課題は、(1)検索対象からの特徴量の抽出法、(2)各検索システムにおけるスコアの計算法、(3)スコアの正規化法、(4)正規化されたスコアの統合法、(5)検索結果の揭示法、の5点に集約されるとし、そのうち、(3)と(4)を研究対象としており、本研究の対象と近いものとなっている。正規化については、まずstandardによりスコアを0から1の範囲に値を納め、また、情報量を導出し、それらの値を統合することにより、スコアを正規化する。正規化されたスコアの統合について、[5]では、相加平均、相乗平均、調和平均、最小値、最大値、PRO関数などの評価関数を用いて、スコアの統合を行っている。

本研究では、複雑な構造を持つXML文書から類似性の高い構造を発見し、さらに、要素値として、単純な数値や文字列だけでなく、テキストや画像などの異種メディアの類似性を考慮し、それらの特徴量の適切な集約手法を提案することを目的とする。

3. 提案手法

本章ではまず、類似度算出の対象とする部分木の探索手法について示す。続いて、木構造間類似度と要素値間

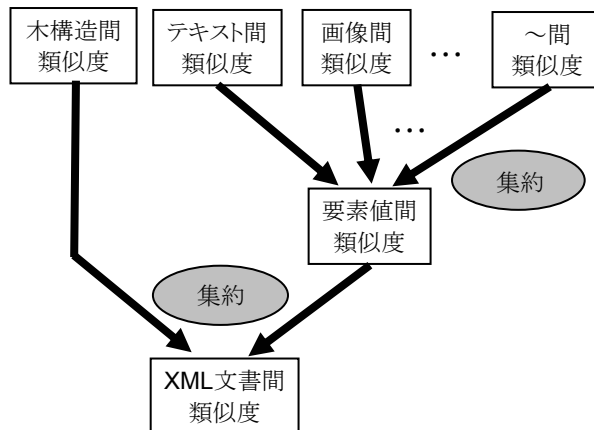


図2 提案手法概念図

類似度それぞれの特徴量の算出法について述べる。その後、導出された2つの異種特徴量の集約手法について述べる。なお、要素値間類似度については、現時点では、テキストと画像の2つのメディアについて、その特徴量を扱うこととする。提案手法の概念図を図2に示す。

3.1 対象部分木の探索

類似度を算出するにあたり、まず、比較対象とする部分木を探索する必要がある。本研究では、比較元の木をテンプレートと呼ぶこととし、そのテンプレートと深さが等しいものを比較対象の部分木とみなすこととする。そこで、次のような手法で対象部分木を算出する。

まず、入力要素について、子要素のリストと数を導出する。次に、もしも子要素がない、つまり、子要素の数が0ならば、処理を終了する。そうでなければ、全ての子要素について、この対象部分木探索アルゴリズムを再帰的に適用する。なお、その前処理として、探索作業の連続上昇回数を1としておく。この連続上昇回数は、テンプレートとの深さの一致の判定に利用する。全ての子要素について処理を終えたあとは、この要素に対する処理が終了し、親要素の処理に戻るため、連続上昇回数を1つ増やす。ここで、もしも連続上昇回数がテンプレートの深さの値と等しいならば、この要素を最上位の親要素とする部分木を比較対象の部分木とし、XML文書間の類似度を算出する。対象部分木の探索アルゴリズムを図3に示す。また、図4に探索の例を示す。この場合、対象部分木の最上位要素はcとjとなる。

3.2 XML構造間類似度の算出法

木構造の類似性を見つけるということは、要素の名前と要素の格納位置が、2つの木の間でどの程度一致しているかを導くことであると考えられる。そこで、XML構造間類似度の算出に際し、Chbeirらが提案する、CBS (Commonality Between Sub-trees) という概念を導入する。

CBSの導出アルゴリズムを図5に示す。CBSは、テンプレートと対象部分木の間で、深さとラベルが一致する要素の個数である。このCBSの値をテンプレートと対象部分木の要素数の平均値で割り、正規化した値をXML構造間の類似度とする。CBSと木構造間類似度の算出例を図6に示す。

3.3 要素値間類似度の算出法

要素値間類似度算出の流れは次のとおりである。まず、算出を行うタイミングであるが、これは、木構造間類似

アルゴリズム 1 : 対象部分木の探索

入力 : Node

```

1 : NodeList children = Node.getChildNode()
2 : Length = |children|
3 : if (Length == 0)
4 :   return
5 : UpNum = 1
6 : for (i=1 ; i<=Length ; i++)
7 :   checkSecondTree(children(i))
8 : UpNum++
9 : if (UpNum == TempDepth)
10 :  XML 文書間類似度算出
11 : Ret XML 文書間類似度

```

図 3 対象部分木探索アルゴリズム

テンプレート	対象木
<a>	<a>
	
<c/>	<e>
	<f/>
	<g/>
	<h/>
	</c>
	<d>
	<i/>
	<j>
	<l/>
	</j>
	</d>
	

図 4 対象部分木の探索例

度算出時の CBS の計算時に、要素同士でタグ名と深さの比較を行うが、ここで同時に、要素値のメディアが一致するか否かを判定し、一致する場合には、その要素値がテキストならばテキスト間類似度、画像ならば画像間類似度と、それぞれ適切な類似度を算出する。ここで導出される全ての類似度はそのメディアの種類とともに保持しておき、CBS 及び木構造間類似度の算出後、後述の 3 つの手法のうちのいずれかを利用して、要素値間類似度の集約を行う。ここで集約された値を、2 つの XML 文書間の要素値間類似度とする。

以下では、まず、本研究で対象とするテキスト間類似度、及び画像間類似度の算出手法について説明する。続いて、要素値間類似度の集約について、3 つの手法を提案する。さらに、要素値間類似度の算出において生じる問題についても説明する。

3.3.1 テキスト間類似度の算出法

本研究では、2 つのテキスト間の類似度算出手法として、次のような一般的な手法を採用する[7]。

まず、それぞれの文書を形態素解析し、名詞語句を抽出する。本実験では、形態素解析ツールとして、Yahoo! デベロッパーネットワークの日本語形態素解析 Web サービスを利用する[8]。次に、抽出された各名詞について TF・IDF を算出する。ここで、TF は、単語 w のテキスト D における頻度、IDF は単語 w を含む文書頻度の逆数である。その結果から、TF・IDF が一定値以上のものを特徴語として採用する。そして、各テキストについてそれらの特徴語が出現するか否かを 1 と 0 で表現し、ベクトルの形にすることで、文書ベクトル d を得る。最後に、そ

アルゴリズム 2 : CBS

入力 : サブツリー $SbTi, SbTj$

```

1 : Dist[][] = new[0...|SbTi|][0...|SbTj|]
2 : Dist[0][0] = 0
3 : For (n=1;n<=|SbTi|;n++)
4 :   Dist[n][0] = Dist[n-1][0] + 1
5 : For (m=1;m<=|SbTj|;m++)
6 :   Dist[0][m] = Dist[0][m-1] + 1
7 : For (n=1;n<=|SbTi|;n++)
8 :   For (m=1;m<=|SbTj|;m++)
9 :     Dist[n][m] = min{
10 :      if (深さとラベル名が一致するならば)
11 :        Dist[n-1][m-1],
12 :        Dist[n-1][m] + 1,
13 :        Dist[n][m-1] + 1 }
14 : Ret (|SbTi| + |SbTj| - Dist[|SbTi|][|SbTj|]) / 2

```

図 5 CBS 導出アルゴリズム

テンプレート	対象部分木	CBS	類似度
			
<c/>	<c/>	3.0	1.0
<d/>	<d/>		
			
	<d>		
	<e/>	0.0	0.0
	<f/>		
	</d>		
			
	<c/>	3.0	0.86
	<d/>		
	<h/>		
			

図 6 CBS と木構造間類似度の算出例

これらの文書ベクトルについて、余弦尺度を導出することにより、2 つのテキスト間の類似度を導出する。余弦尺度は以下の式で与えられる。

$$Sim_{text} = \frac{d_i \cdot d_j}{\|d_i\| \|d_j\|} \quad (1)$$

上記の手法よりも直感的な手法として、2 つの文書の形態素解析により得られる全名詞数に対する重複語句数の割合である、Jaccard 係数を類似度として利用することもできる。

3.3.2 画像間類似度の算出法

画像間類似度の算出については、さまざまな手法が提案されているが、画像処理では、比較的大きい計算量のものもある。本研究は画像間類似度を高精度に求めることを主の目的としないので、以下のような直感的かつ簡易的な手法で、画像間類似度を算出する。

まず、各々の画像を、RGB それぞれ 256 色で表現されたラスタ画像の形式で用意する。それらの画像について、各々ヒストグラムを算出し、ピクセル数で割って正規化する。そして、算出された 2 つのヒストグラムについて、マンハッタン距離を導出する。マンハッタン距離は以下の式で与えられる。

$$d_{ij} = \sum_{k=1}^m |x_{ik} - x_{jk}| \quad (2)$$

この値の逆数を、画像間の類似度とする。

上記の類似度は、形状など、色頻度以外に画像の特徴となる情報を考慮していないため、画像間の全体的な色あいの類似性のみを判定することになる。

3.3.3 要素値間類似度の集約法

構造間類似度との集約に向けて、まず、導出した全ての要素値間類似度を集約し、統一した値を求める。要素間類似度の集約手法として、3つの手法を提案する。

1つ目の手法は、導出した全ての要素値間類似度を合計し、テンプレートと対象部分木のタグ値の組み合わせの数で割ることにより平均を導出して、その値を集約された要素値間類似度とするというものである。この手法では、単純にテンプレートと対象部分木のタグ値の組み合わせの数で割るので、タグ値のメディアが一致しない場合の要素間類似度を0と考慮して集約を行うことになる。

2つ目の手法は、導出した全ての要素値間類似度を合計し、要素値のメディアが一致した数、すなわち、要素値間類似度の数で割ることにより平均を導出して、その値を集約された要素値間類似度とするというものである。この手法では、メディアが一致しない場合の要素間類似度は考慮せずに集約することになる。

3つ目の手法では、まず、テキストと画像それぞれについて、要素値間類似度を合計し、各々の類似度の数で割ることにより平均を導出する。そして、求められた2つの値について、次の式で示される重み付き平均を導出し、その値を集約された要素値間類似度とする。

$$Sim_{value} = \alpha \times Sim_{text} + \beta \times Sim_{image} \quad (3)$$

(ただし、 $\alpha + \beta = 1$)

このとき、いずれかの要素値間類似度が存在しない場合、もう1種類の要素値間類似度の値を集約された要素値間類似度とする。

前者2つの手法では、メディアの種類を区別せずに集約を行うという特徴を持ち、後者の手法はメディアの種類を区別して集約を行うという特徴を持つ。

テンプレート
<a> img1.jpg <c>本日は晴天なり</c>

対象部分木	手法毎 XML 文書間類似度		
	1	2	3
<a> img1.jpg <c>本日は晴天なり</c> 	0.75	1.0	1.0
<a> img1.jpg <c>本日は晴天なり</c> <d>img2.jpg</d> 	0.62	0.82	0.85
<a> img1.jpg <c>本日は晴天なり</c> <d>静岡県浜松市</d> 	0.60	0.76	0.80

図7 要素値間類似度算出手法の違いによるXML 文書間類似度の変化例

テンプレート	要素値数	XML 文書間類似度
<a>	1	1.0
img1.jpg	2	1.0
<c>本日は晴天なり</c>	3	0.92
<d>img2.jpg</d>	4	0.79
<e>静岡県浜松市</e>	5	0.77
<f>img3.jpg</f>	6	0.73
<g>2008年3月</g>	7	0.71
<h>img4.jpg</h>	8	0.70
...		
		

図8 要素値数の変化に伴うXML 文書間類似度の低下例

また、上記3つのいずれの手法にも共通する問題として、それぞれの要素値間類似度が、メディアが一致するか否かという条件のみで導出されるか否かが決定されているため、要素値の数が多くなるに従って、要素値間類似度の数が多くなり、結果として、集約された要素値間類似度の値が小さくなってしまおうという点が上げられる。木構造間類似度は、テンプレートと対象部分木が完全に一致する場合にはその値が1となるが、要素値間類似度は、完全に一致する場合でも、ほとんどの場合、その値が1になることはない。要素値間類似度の集約法の違いによるXML 文書間類似度の変化を図7に示す。なお、以下の実験で利用した画像の一覧は、巻末の付録に掲載する。

3.4 木構造間類似度と要素値間類似度の集約法

算出された構造間類似度と要素値間類似度の集約を行い、XML 文書間の類似度を求める。類似度を集約する手法として、本研究では要素値間類似度集約の場合と同様に、構造間類似度と要素間類似度の重み付き平均を導出し、その値をXML 文書間の類似度とする。

ただし、ここで考慮しなくてはならない点が、3.3.3節で述べた要素値間類似度に関する問題である。要素値間類似度は、2つのXML 文書が完全に一致する場合でも、1となるのが稀であるため、木構造間類似度との間で集約を行うにあたり、計算手法を工夫しなくてはならないと考えられる。本研究では、同一のテンプレートに対して導出された類似度同士を比較し、順序付けすることにより検討を行うので、数値そのものの妥当性については言及しないこととする。要素値数の違いによるXML 文書間類似度の変化の例を図8に示す。要素値間類似度の集約手法は、手法3を用いる。また、対象部分木はテンプレートと同一のものを利用する。

4. 予備実験

本章では提案手法の妥当性を検証するために行った実験について述べる。

4.1 実験データ

本実験では、提案手法が正常に機能していることを確かめるために必要な性質を持つ十分な量のデータを準備しようと考え、既存のXML データではなく、個人情報ジェネレータである“なんちゃって個人情報”[9]により生成したデータを、テキストエディタにより編集したものを用いた。具体的には、構造については、各部分木内の要素値を持つ要素は1個から6個までであり、かつ、これらの要素の組み合わせそれぞれについて100件ずつ、

合計 6300 件の部分木を対象とした。また、要素値については、テキストであるものが 3 つ、画像であるものが 3 つで、テキストと画像の各々で、それぞれの要素に 2 種類、4 種類、6 種類の要素値が格納されている。要素値数が 6 個である部分木の例を図 9 に示す。例として、要素 animal と要素 snow に格納される要素値はそれぞれ 2 種類である。なお、この木を本実験のテンプレートとして用いた。

4.2 実験結果

実験データ中の 6300 個の全ての部分木について、まず、テキスト間類似度と画像間類似度の重みをともに 0.5、木構造間類似度と要素値間類似度の重みをともに 0.5 として、提案手法を実装したプログラムにより図 9 のテンプレートとの間で類似度を算出した。その結果、図 10 から図 12 に示す各部分木が、テンプレートとの類似性が高

```
<record>
  <animal>img01.jpg</animal>
  <place>
    静岡県は東海地方にあり太平洋に面する
  </place>
  <snow>
    太平洋沿岸は雪が少ない
  </snow>
  <neighbor>
    北海道は他の地方と接していない
  </neighbor>
  <player>img001.jpg</player>
  <flower>img1.jpg</flower>
</record>
```

図 9 実験用テンプレート

```
<record>
  <animal>img01.jpg</animal>
  <place>
    静岡県は東海地方にあり太平洋に面する
  </place>
  <snow>
    太平洋沿岸は雪が少ない
  </snow>
  <neighbor>
    北海道は他の地方と接していない
  </neighbor>
  <player>img001.jpg</player>
  <flower>img2.jpg</flower>
</record>
```

図 10 類似度上位部分木(1)

```
<record>
  <animal>img01.jpg</animal>
  <place>
    宮城県は東北地方にあり太平洋に面する
  </place>
  <snow>
    太平洋沿岸は雪が少ない
  </snow>
  <neighbor>
    北海道は他の地方と接していない
  </neighbor>
  <player>img005.jpg</player>
  <flower>img1.jpg</flower>
</record>
```

図 11 類似度上位部分木(2)

```
<record>
  <animal>img01.jpg</animal>
  <place>
    静岡県は東海地方にあり太平洋に面する
  </place>
  <snow>
    太平洋沿岸は雪が少ない
  </snow>
  <neighbor>
    中国地方は関西地方と接する
  </neighbor>
  <player>img001.jpg</player>
  <flower>img1.jpg</flower>
</record>
```

図 12 類似度上位部分木(3)

表 1 類似度上位部分木との各類似度

順位	木構造間	要素値間	XML 文書間
1	1.0	0.449	0.725
2	1.0	0.42949	0.71475
3	1.0	0.42943	0.71471

```
<record>
  <animal>img01.jpg</animal>
  <player>img001.jpg</player>
  <flower>img1.jpg</flower>
</record>
```

図 13 類似度上位部分木(4)

いと判定された。それぞれの木構造間類似度、要素値間類似度、及び XML 文書間類似度を表 1 に示す。

次に、テキスト間類似度と画像間類似度の重みを 0.2 : 0.8 と 0.8 : 0.2 の 2 パターンに設定し、それぞれテンプレートと実験データとの類似度を算出した。その結果、テキスト間類似度と画像間類似度の重みが 0.2 : 0.8 の場合は図 12 の部分木が、0.8 : 0.2 の場合には、図 10 の部分木がそれぞれ最も類似度が高いと判定された。

さらに、木構造間類似度と要素値間類似度の重みを 0.2 : 0.8 と 0.8 : 0.2 の 2 パターンに設定し、それぞれテンプレートと実験データとの類似度を算出した。その結果、木構造間類似度と要素値間類似度の重みが 0.8 : 0.2 の場合は図 10 の部分木が、0.2 : 0.8 の場合には、図 13 の部分木がそれぞれ最も類似度が高いと判定された。

4.3 考察

図 9 のテンプレートと、類似度上位であった図 10、図 11、図 12 の対象部分木を比較すると、目視ではあるが、ほぼ類似したデータとなっており、良好な結果が得られたと考える。

まず、対象部分木同士の順位関係に着目する。はじめに、図 10 のデータが図 11 のデータよりも上位である点について考えると、これは、図 10 のデータが flower 要素の要素値のみテンプレートと異なっているのに対し、図 11 のデータは place 要素の要素値と player 要素の要素値の 2 つが異なっているためであると考えられる。次に、ともに 1 つの要素値だけが異なっているにもかかわらず、図 10 のデータが図 12 のデータよりも上位である点について考えると、これは、一致していない要素値間の類似度について、図 10 で異なっている画像間類似度の

ほうが、図 12 で異なっているテキスト間類似度よりも高い値となったためであると考えられる。最後に、図 11 のデータのほうが図 12 のデータよりもテンプレートとの間で異なる要素値数が多いにもかかわらず、図 11 のデータが図 12 のデータよりも上位である点について考えると、これも、上記の場合と同様に、図 11 において異なっているテキスト、画像の類似度が、図 12 において異なっているテキストの類似度よりも高い値であったためであると考えられる。

次に、テキスト間類似度と画像間類似度の重みを変化させた場合の結果に着目する。まず、画像間類似度の重みを高くした場合には、図 12 の部分木との類似度が 1 番高くなっているが、これは、画像要素値は完全に一致しており、また、テキスト要素値の不一致も 1 つだけであるためであると考えられる。また、テキスト間類似度の重みを高くした場合には、図 10 の部分木との類似度が 1 番高くなっているが、これは、テキスト要素値は完全に一致しており、また、画像要素値の不一致も 1 つだけであるためであると考えられる。

最後に、木構造間類似度と要素値間類似度の重みを変化させた場合の結果に着目する。まず、木構造間類似度の重みを高くした場合には、図 10 の部分木との類似度が 1 番高くなっているが、これは、要素値が 6 つの部分木との木構造間類似度は全て 1.0 であり最も高いので、全ての重みを 0.5 とした場合との順位の変化は起こらないためであると考えられる。また、要素値間類似度の重みを高くした場合には、図 13 の部分木との類似度が 1 番高くなっているが、これは、例えば図 12 のように木構造と画像要素値がテンプレートと一致する部分木が、テキスト間類似度により要素値間類似度が低下し、図 13 の要素値間類似度を下まわってしまうためであると考えられる。この点は今後改善しなくてはならない点であると考えられる。

5. おわりに

本研究では、XML 文書に関して、木構造の類似度、及び複数種メディアからなる要素値の類似度という異なる特徴量を集約することにより、高精度に類似性を発見する手法を提案した。

今後の課題として、テキストと画像以外のメディアの特徴量の導出と集約を行い、提案手法についてさらに深い評価と検討を行う必要があると考えられる。本文中で述べている、要素値数の増加による類似度低下という問題の解決法について検討する必要がある。

類似度の集約手法については、本論文では重みを変化させた場合に XML 文書間類似度がどのように変化するかということのみ示した。今後は、それぞれの類似度が持つ情報量を考慮した上で、適切な重みを見つけ出すことが課題である。考察で述べたように、一方の類似度が集約した類似度を低下させてしまうという問題がある。

現在は、比較元の XML パターンとしてあらかじめテンプレートを用意しているが、より実用的なものに近づけるためには、テンプレートとして用いる特徴的な XML パターンの抽出手法についての検討が考えられる。その際にも、構造と要素値の内容という 2 つの視点から特徴的であると考えられる部分木を抽出する必要がある。

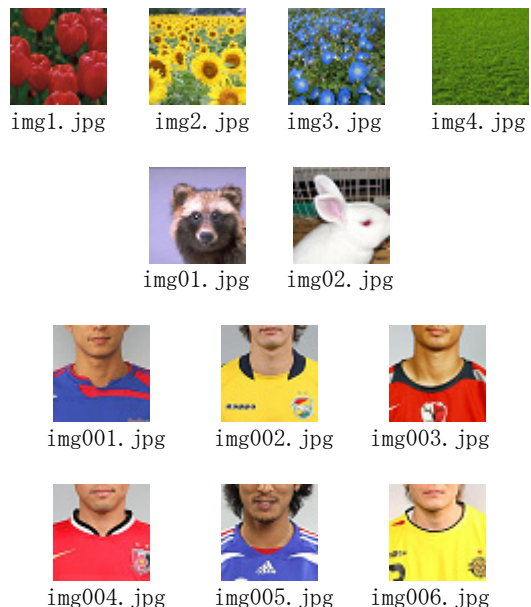
参考文献

[1] microformats, <http://microformats.org/>.

- [2] Y. Ma and R. Chbeir, “Content and Structure Based Approach for XML Similarity”, In Proc. The Fifth International Conference on CIT, Dallas, US, October. 2005.
- [3] 文連子, 天笠俊之, 北川博之, “木直列化を用いた XML データの類似結合”, 電子情報通信学会技術研究報告, Vol.107, No131, pp.91-96, 電子情報通信学会 (2007).
- [4] X. Gong, W. Qian, Y. Yan, and A. Zhou, “Bloom Filter-based XML Packets Filtering for Millions of path Queries”, Proceedings of the 21st International Conference on Data Engineering (ICDE), 2005.
- [5] 鈴木優, 波多野賢治, 吉川正俊, 植村俊亮, “複数のメディアで構成された電子文書の検索手法”, 情報処理学会論文誌, データベース, Vol.42, No. SIG 10 (TOD 11), pp.11-21 (2001).
- [6] 鈴木優, 波多野賢治, 吉川正俊, 植村俊亮, “検索結果を統合するための情報量の概念を考慮したスコア正規化手法”, 情報処理学会論文誌, データベース, Vol.45, No. SIG 4 (TOD 21), pp.37-49 (2004).
- [7] 石川博, “次世代データベースとデータマイニング”, pp.182-187, CQ 出版株式会社, 東京, 2005.
- [8] Yahoo!デベロッパーネットワーク, <http://developer.yahoo.co.jp/>.
- [9] なんちゃって個人情報, <http://kazina.com/dummy/index.html>.

付録

本研究の実験に利用した画像を以下に示す。



謝辞

本研究の一部は科学研究費補助金基盤研究(B)(課題番号 19300026)の助成による。