Web コンテンツにおけるオブジェクトの準同一性判定

平野　雄也† 　馬　　強† 　吉川　正俊†

† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町

E-mail: † y.hirano@db.soc.i.kyoto-u.ac.jp, {qiang,yoshikawa}@i.kyoto-u.ac.jp

あらまし　Web 上には，過去から現在に至るまでの膨大な情報が蓄積されている．こうした情報の中には，同一オブジェクト（組織，人物など）に関する異なる記述が多数存在する．そこで本研究では，異なるページ間における同一オブジェクトの差分や推移の分析を支援するため，Web ページから条件付確率場（Conditional Random Fields，CRF）を利用してオブジェクトに関する記述を抽出し，構造化する．そして，更新時間を考慮して準同一関係にあるオブジェクトを発見する手法を提案する．大学の Web サイトを対象にして研究室の組織などの準同一性を判定する実験を行い，本手法の有効性を検証する．

キーワード　準同一関係，情報抽出，Web データベース

1. はじめに

近年，Web の発展に伴い，過去から現在に至るまでの膨大な情報が Web 上に蓄積されている．こうして蓄積された情報の中には，同一オブジェクト（組織，人物など）に関する異なる記述が多数存在する．記述の不一致は，大きく次の 2 つに分類される．

- オブジェクト推移（バージョンアップ）に伴う記述の更新による，時間軸での不一致
- 記述の更新タイミングや情報量の違いによる，サイト間での不一致

ここで，Web サイト s_1, s_2 において，オブジェクト O が図 1 のように推移した場合を例にとり説明する．時刻 t_1, t_2 における，オブジェクト O, O' の違いは，時間軸での不一致を表している．また，時刻 t_3 における，サイト s_1 のオブジェクト O'' とサイト s_2 のオブジェクト O' の違いは，サイト間での不一致を表している．

実際の Web サイトを例に考えると，大学や企業などの組織では，各部局の Web サイトによって情報の公開が行われることが多い．京都大学の場合，大学，学部，学科，大学院研究科，専攻，研究室など，多くのサイトが互いに関連し合い，情報の公開を行っている．こうした複数のサイト間では，教員情報，研究テーマ，発表論文などが複数のページに記述され，サイトの管理者やページの更新時期が異なるために記述の不一致が起こりやすいという問題がある（図 2）．このような不一致を修正するには，膨大なコストがかかる．特に，組織が大規模になるほどサイトも大規模になり，不一致の記述箇所を発見するには多大な労力を必要とする．以上の背景から，次のようなニーズが存在すると考えられる．

- 大規模サイトの管理・保守支援
  - 同一オブジェクトに関する差分検出
  - 他サイトのオブジェクト変更通知
- オブジェクトの遷移分析
  - 過去と現在の情報を比較

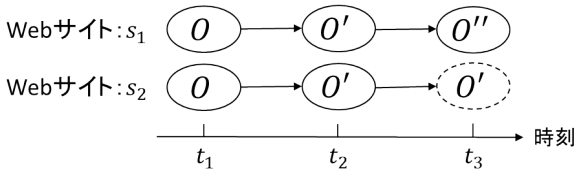

図 1　オブジェクト推移の例

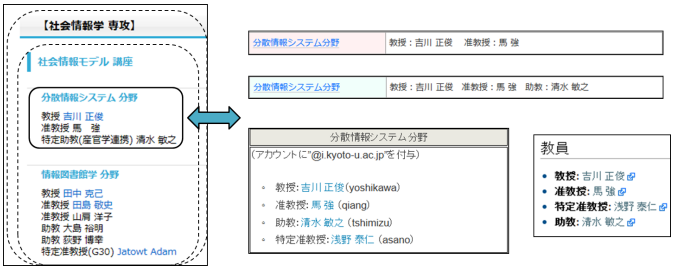

図 2　準同一関係にあるオブジェクトの例

そこで本研究では，複数のページ間で準同一関係にあるオブジェクトを発見する手法を開発し，準同一オブジェクトの差分や推移の分析を支援することを目的とする．全体の流れは図 3 のようになる．オブジェクトに関する記述を，関連するデータ（Web ページ中に出現する単語）の集合と定義する．このデータは，オブジェクトの特定の属性に対する属性値ともいえる．Web ページ中に記述されるオブジェクトは，図 2 左に示すように，階層構造を成していることが多い．また，オブジェクトは複数のレコードの属性値の集合から構成される．そこで，Web コンテンツのオブジェクトを木構造によってモデル化する．これにより，オブジェクトの階層関係と，各レコードの構成要素を表現できる．オブジェクトに関する記述は，Web ページ上では表やリストなどによって表現されることが多いため，このような半構造データを対象に抽出を行う．Web コンテンツにおけるオブジェクトに関する記述を木構造によりモデル化したものを，オブジェクト木と呼ぶ．

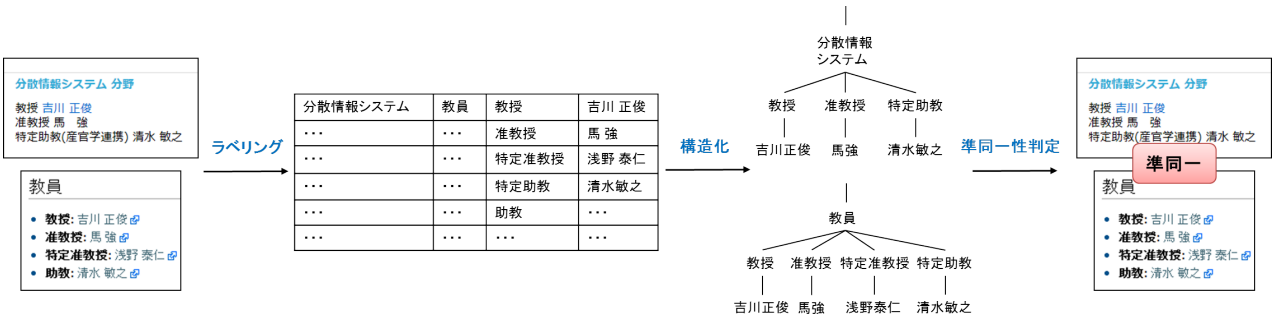

図 3　準同一オブジェクト発見の流れ

本論文では，オブジェクトの記述を表現したオブジェクト木がきわめて類似するとき，準同一関係が成立するという．すなわち，記述間の類似度が閾値以上のものを，準同一オブジェクト対として検出する．閾値は，ページ間の更新時間差によって定める．更新時間の差によって，異なるページ間の同一オブジェクト同士で記述の不一致が発生する可能性が高くなるため，その違いを許容するために用いる．オブジェクトの違いを許容するために，準同一という概念を導入したが，これは，準同一映像を検出する研究 [1]（NDD, Near Duplicate Detectator）で使用される用語である．

本論文の構成は以下の通りである．第 2 節では，関連研究について紹介する．第 3 節では，Web ページからオブジェクトに関する記述を抽出し，構造化したのち，オブジェクト間に準同一関係が成立するかどうかを判定する手法を提案する．第 4 節では，提案手法の有効性を検証するために，大学の Web サイトを対象に評価実験を行う．最後に第 5 節では，本論文をまとめ，今後の課題について述べる．

2. 関 連 研 究

2.1 準同一映像の検出（NDD）

準同一映像とは，映像中に複数回出現する，きわめて類似した画像特徴を持つ映像断片のことである．準同一映像の検出は，Near Duplicate Detection（NDD）ともいわれる．NDD では，区間の照合の計算量を減らすために特徴量の圧縮などが行われる．本研究で用いる “準同一” という用語は，NDD に関する研究で用いられている．社本ら [1] は，準同一区間の出現パターンを分析することで，準同一映像を再放送，広告，タイトル，予告などに分類する手法を提案している．

2.2 Web ページからの情報抽出

Web ページから構造化データを自動抽出する手法の研究はこれまでに多く行われてきた．情報抽出の手法として，自然言語処理を用いるもの [2]，Web ページの構造を用いるもの [3]，これらを組み合わせたもの，などがある．

自然言語処理を用いる手法 [2] では，テキストの文構造に注目する．コーパスから単語間の関係を抽出する．例えば，“person” is author of “booktitle” という関係から，本のタイトルと著者が抽出される．Web ページの構造を用いる手法 [3] では，DOM 木の構造に注目する．DOM 木から，繰り返し現れるテンプレートを発見し，その部分からデータ抽出を行うというものである．また，Zheng ら [4] は，DOM 木中の各ノードに含まれるラベリングされた葉ノードの平均情報量（エントロピー）を求め，データリッチなノードを特定することでデータ抽出を行う手法を提案している．

2.3 Web コンテンツの制約発見

本研究の目的に近い研究として，Web コンテンツの包含関係の発見を行う研究がある [5] [6]．Web ページにおいて成り立つ包含関係として例えば，研究室の各構成員の発表論文一覧は，研究室の発表論文一覧のサブセットである．

澤ら [5] は，HTML 文書中の繰り返し構造（リストなど）に注目し，Web ページのコンテンツを木構造によりモデル化し，タグパスで指定されるノード集合の包含関係を発見する手法を提案している．高橋ら [6] は，Web ページのブロック内に含まれる単語集合を比較することで，コンテンツの包含関係を発見する手法を提案している．

それに対し本研究では，オブジェクトに関する記述を抽出・構造化し，データの意味を解釈することでオブジェクト同士で粒度の細かい比較を行うことができ，差分や推移の分析が可能となる．

2.4 重複データの検出

企業では，顧客情報などが複数のデータベースに保管されるため，それらのデータを統合したいというニーズがある．文献 [7] では，データベース（RDB）から重複するレコードを発見する手法が紹介されている．本研究で扱うオブジェクト木は，データベースのレコード集合に相当する．また，オブジェクトを階層的に表現できる．本研究では，データの構造化をも対象とする．

3. 提 案 手 法

準同一関係にあるオブジェクトを発見するために，Web ページの記述からオブジェクトに関する属性情報を抽出し，構造化した後，準同一性を判定する．この章では，これらの手法について順に説明する．

3.1 構造化データの抽出

Web ページの記述からオブジェクトに関する属性情報を抽出

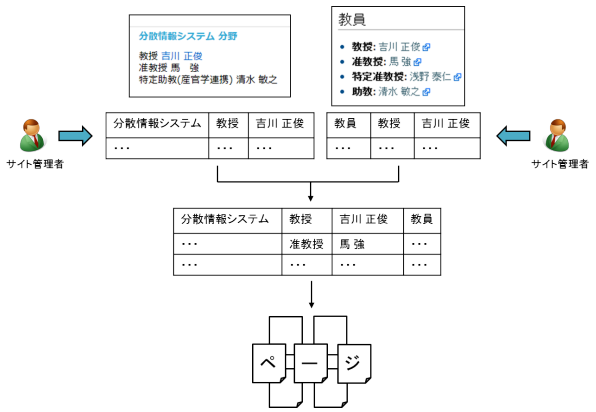

図 4　CRF ラベリングの手順

<h3> 社会情報学専攻 </h3>

<th> 分散情報システム分野 </th>

<td> 教授 : 吉川 正俊　准教授 : 馬 強　助教 : 清水 敏之 </td>

“専攻名”	“職名”	“氏名”
社会情報学専攻	教授	吉川 正俊

図 5　ソースコードと属性値集合の例

するために，統計的機械学習に基づく手法である CRF [8] を利用し，系列ラベリング問題として定式化する．すなわち，Web ページのソースコードを，単語や記号，HTML タグなどの要素が連なった系列とみなし，各要素に対してラベル付与することを行う．テキストノード中の各単語は，スペースと記号によって分割する．オブジェクトの属性名がラベルとなり，各単語がどの属性に分類されるかを考える．

文献 [9] によれば，系列を $\boldsymbol{x}$，ラベル列を $\boldsymbol{y}$ と表すと，CRF は対数線形モデルの一種であるため，その条件付き確率 $P(\boldsymbol{y}|\boldsymbol{x})$ は，

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{x},\boldsymbol{w})} \cdot \exp(\boldsymbol{w} \cdot \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}))$$

と表される．ただし，$\boldsymbol{f}(\boldsymbol{x},\boldsymbol{y})$ は素性関数ベクトル，$\boldsymbol{w}$ は素性に対する重みベクトルである．また，$Z(\boldsymbol{x},\boldsymbol{w})$ は正規化項であり，$Z(\boldsymbol{x},\boldsymbol{w}) = \sum_{\boldsymbol{y}} \exp(\boldsymbol{w} \cdot \boldsymbol{f}(\boldsymbol{x},\boldsymbol{y}))$ と定義される．CRF は，学習データをもとに重み $\boldsymbol{w}$ の学習を行い，確率 $P(\boldsymbol{y}|\boldsymbol{x})$ を最大にするラベル列を推定する．

本研究では，図 4 に示すような状況を想定して各ページに対してラベリングを行う．まず，サイトの管理者が，あるページのいくつかのデータに対して，抽出対象となるオブジェクトの属性値集合を指定する．これを教師データとして，そのページの残りのデータに対して，CRF を適用し，自動でラベリングを行う．このようにして得られた各ページのラベリング結果を統合することで，大きな一つの学習データを得る．これをもとに，すべてのページに対してラベリングを行う．

Web ページから構造化データを抽出するために CRF を用いる利点として，次のようのものが挙げられる．

- 前に出現した文字列のラベリング結果を素性とできる
- テキスト構造と DOM 構造を考慮したデータの抽出

そこで，CRF の素性は，具体的に次のようなものを用いる．

- 文字列自体の特徴（文字列 x がラベル y に割り当てられる）
- ラベル間のつながり（前文字列のラベル y' であるとき，文字列 x がラベル y に割り当てられる）

図 5 のような例を考えると，素性として「th タグの直後にラベル “専攻名” が割り当てられる」,「ラベル “職名” の直後にラベル “氏名” が割り当てられる」などが用いられ，それらの素性に対する重みは大きい．

3.2　Web コンテンツの構造化

CRF を利用してラベリングされたデータ集合をもとに，Web コンテンツをモデル化したオブジェクト木を構築する．ここでは，関連のあるデータ集合，すなわち，あるオブジェクトの構成要素となるデータ集合を特定する手法を述べる．そのために，Web ページの DOM 構造における各ラベルの出現順序と出現頻度を利用する．これは，以下の考察に基づく．

- 表やリストなどを含む Web ページではオブジェクトが階層的に分類されて表示されることが多いため，各ラベルの出現頻度は異なる．出現頻度が小さいラベルほど，より上位階層のオブジェクトのデータとなる．
- オブジェクトが階層的に分類される場合，より上位階層のオブジェクトのデータは DOM 構造における上もしくは左の位置に存在する．
- オブジェクトを構成する各データは，他のオブジェクトを構成するデータよりも近い位置に存在する．

これらの考察に基づき，DOM 構造とラベルの出現頻度から，データを階層化したオブジェクト木を構築する．各データにラベル付与された DOM 木に対して，プレオーダーで探索をはじめ，直前に出現したデータに付与されたラベルの出現頻度と比較することで，新たなデータのノード挿入位置を決定する．

3.3　オブジェクトの準同一性判定

オブジェクトの準同一性を判定するために，構築されたオブジェクト木の類似度を計算する．木の類似度計算には，木編集距離 [10] がよく利用される．木の編集距離とは，二つの木を同じ木に変換するための編集操作の最小のコストである．編集操作には，ノードの挿入，削除，置換の三つがある．木の編集距離が小さいほど，木の類似度は大きくなる．この手法は，順序木を対象としているが，本研究で対象とするオブジェクト木は兄弟の順序に意味を持たない非順序木である．さらに，親子の階層の順序にも意味を持たない．オブジェクト木は，図 6 のように，各葉ノードから根ノードまでの経路上にあるノード集合をひとつのレコードとみなしたレコード集合と考えることができる．このレコード間の対応付けを行うことにより，オブジェクト木におけるノードの対応付けを考える．

以上の議論から，次式により二つのオブジェクト木 T_A と T_B の類似度 $Similarity(T_A, T_B)$ をオブジェクト木の類似度を定義する．この類似度が閾値以上であるとき，準同一オブジェクトであると判定する．

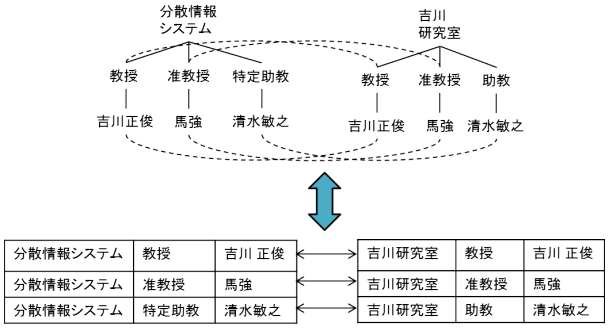

図 6　オブジェクト木とレコード集合における対応付け

$$Similarity(T_A, T_B) = \frac{2 \cdot MatchNodes(T_A, T_B)}{Nodes(T_A) + Nodes(T_B)} \quad (1)$$

ここで，

- $Nodes(T_A)$，$Nodes(T_B)$ は，それぞれ T_A，T_B のノード数を表す
- $MatchNodes(T_A, T_B)$ は，T_A，T_B において一致したノード数を表す

とする．提案手法では，より葉に近い小さな部分木同士の類似度を先に計算する．これは，二つのオブジェクト木が準同一であるとき，準同一となる部分木が必ず存在するためである．

次に，閾値をページの更新時間差により動的に決定する方法を述べる．Web ページ間で記述の不一致が起こる要因の一つとして，更新時間の違いがある．例えば，同じような情報を記述した二つのページのうち，一方の更新時間が 1 年以上も前で，もう一方のページの更新時間が最近のものであり二つのページ間で記述の不一致があれば，1 年以上も更新を行われていないページの情報は古いものであり誤っていると推測される．そのため，二つのページ間の更新時間の差を，相当関係を計算する際に考慮する必要がある．二つのページの更新時間の差が小さければ，オブジェクト木のわずかな違いでも，対象コンテンツが異なるオブジェクトを表している可能性が高い．つまり，準同一でないと判断する．一方，この二つのページの更新時間の差が大きければ，更新時期の違いにより記述の違いが生じた可能性が高いため，オブジェクト木の違いがやや大きくても，準同一である可能性がある．

ページ A，B の更新時間を $time(A)$，$time(B)$ と表すことにすると，閾値をその更新時間の差によって定める．更新時間の差が大きいほど，オブジェクトの記述間の差異も大きくなる可能性が高いため，閾値が小さくなるようにする．また，閾値は上限と下限を設定する．これは，閾値が 1 に近すぎる場合，準同一オブジェクトの発見が困難になってしまい，また閾値が 0 に近すぎる場合には準同一でないオブジェクトを発見してしまうためである．ページ A，B の更新時間 $time(A)$，$time(B)$ によって閾値 $TimeDifference(A, B)$ を次の式で定義する．

$$TimeDifference(A, B) = \begin{cases} \lambda \cdot \mathrm{e}^{-\frac{|time(A)-time(B)|}{\mu}} & (|time(A) - time(B)| < \mu) \\ \frac{\lambda}{\mathrm{e}} & (\text{それ以外}) \end{cases} \quad (2)$$

ただし，λ と μ はともにパラメータとする．λ は閾値の上限であり，μ は指数関数の逓減率とし，閾値の下限をとるときの更新時間の差とする．

4. 実　　験

大学の Web サイトを対象にして提案手法を適用し，研究室の組織や人物などの準同一性を判定する実験を行い，本手法の有効性を検証する予定である．

5. おわりに

本論文では，準同一関係にあるオブジェクトを発見するために，Web ページの記述からオブジェクトに関する属性情報を抽出・構造化し，準同一性を判定する手法を提案した．属性情報の抽出には，CRF を利用し，DOM 構造，自然言語を含む半構造データである Web ページからの抽出を可能にした．ラベリング結果を利用することで，オブジェクトを階層的に表現した木構造でモデル化した．オブジェクト木の類似度を，更新時間を考慮して計算することで，準同一判定を行った．今後の課題として，実験により手法の有効性を検証する必要がある．

文　　献

[1] 社本裕司, 井手一郎, 出口大輔, 高橋友和, 村瀬洋, “大規模放送映像アーカイブにおける出現パターンによる準同一映像区間の分類”, 情報処理学会論文誌, Vol.52, No.12, pp3593-3598, 2011.

[2] Michele Banko, Michael J Cafarella, Stephen Soderland, Matt Broadhead, Oren Etzioni, “Open Information Extraction from the Web”, Proceedings of the International Joint Conferences on Artificial Intelligence, pp.2670-2676, 2007.

[3] Bing Liu, “Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data”, Springer-Verlag, chapter9-10, pp.323-458, 2011.

[4] Xiaoqing Zheng, Yiling Gu, Yinsheng Li, “Web data extraction based on partial tree alignment”, Proceedings of 20th International World WideWeb Conference (WWW2012), pp.93-102, 2012.

[5] 澤菜津美, 森嶋厚行, 飯田敏成, 杉本重雄, 北川博之, “コンテンツ一貫性制約を用いた Web サイト管理手法の提案”, 電子情報通信学会第 18 回データ工学ワークショップ (DEWS2007) 論文集, 2007

[6] 高橋公海, 森嶋厚行, 松本亜季子, 杉本重雄, 北川博之, “Web コンテンツ管理のための一貫性制約発見手法”, DBSJ Journal, Vol.7, No.3, pp.25-30, 2008.

[7] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios, “Duplicate Record Detection: A Survey”, IEEE Transactions on Knowledge and Data Engineering, Vol.19, No.1, pp.1-16, 2007.

[8] John Lafferty, Andrew McCallum, Fernando Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data”, Proceedings of the International Conference on Machine Learning (ICML-2001), 2001.

[9] 高村大也, 奥村学, “言語処理のための機械学習入門”, コロナ社,

2010.

[10] Philip Bille, “A Survey on Tree Edit Distance and Related Problems”, Theoretical Computer Science, Vol. 337, pp. 217-239, 2005.