社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143（82）
2007／7／4

属性値が時間変化するWebオブジェクトの識別・検索手法の提案

白砂 健一† 小山 聡† 田中 克己†

† 京都大学大学院情報学研究科 社会情報学専攻
〒 606-8501 京都府京都市左京区吉田本町
E-mail: †{shirasuna,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

あらまし オブジェクト識別・検索の既存の研究においては，属性値が時間的に変化するオブジェクトを扱うことが困難であった．本稿では，Webから抽出したオブジェクトの属性値に対して，その値の有効時間を検出する方法を提案する．また，属性値の時間変化を考慮したオブジェクト識別方式，および属性値の有効時間を反映したオブジェクト検索結果の呈示方式についても検討を行う．

キーワード オブジェクト検索，オブジェクト識別，クラスタリング，データマイニング

Identification and Retrieval of Web Objects with Time-Varying Attribute Values

Kenichi SHIRASUNA†, Satoshi OYAMA†, and Katsumi TANAKA†

† Graduate School of Infomatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto, 606-8501 Japan
E-mail: †{shirasuna,oyama,tanaka}@dl.kuis.kyoto-u.ac.jp

Abstract In the current reserch of object identification and object search, it is difficult to handle attributes which vary by time. In this paper, we propose a method how to extract valid times of attibute values of objects extracted from WWW. We also propose an object identification method using changes of attibute values, and how to display the result of object search which takes into account the valid times of attribute values.

Key words object search, object identification, clustering, data mining

1. はじめに

既存のWeb検索エンジンの多くは，ユーザからキーワードを受け取り，キーワードを含むページのリストを返す方式を採っている．ユーザの入力するキーワードには様々なものがあるが，その中でも人物や商品，組織や地域など，実世界のオブジェクトに関するキーワード，特にオブジェクト名をキーワードとして用いた検索は大きな割合を占める．例えば，文献[1]によれば，人名をキーワードとした検索はWeb検索の5-10%を占めるとされる．

実際にユーザが必要とするのは，ある特定のオブジェクトに関する情報であるが，ページ単位でしか検索結果を返さない既存の検索エンジンを用いた場合，対象のオブジェクトの情報が複数のページに分散していたり，同一のページ内に対象オブジェクトと無関係な情報が混在していたりする．そのため，目的の情報を獲得するために大きな労力を必要とする場合が多い．そこで，このような問題を解決するため，オブジェクト単位で情報を収集，集約して呈示するオブジェクトベースの検索エンジン[2][3]を構築するための研究が行われてきている．

既存の研究[2]においては，学術論文のように，属性値（著者名や掲載雑誌名）が時間的に変化しないオブジェクトを主に扱ってきた．しかし，人物や企業のようなオブジェクトでは，属性値（職業や年齢，社長や資本金額など）が時間に従って変化する．情報が上書きで更新される管理されたデータベースと異なり，Webにおいては過去の情報が残ったまま，新しい情報が追加されることが多い．そのため，属性値が変化した同一オブジェクトの情報が，別のオブジェクトの情報と誤って判定されたり，属性値が矛盾しているように見えたりといった問題が生じる．

そこで本稿では，Webから抽出したオブジェクトの属性値に対して，その値の有効時間（Valid Time）を検出する方法を提案する．また，属性値の時間変化を考慮したオブジェクト識別方式，および属性値の有効時間を反映したオブジェクト検索結果の呈示方式についても検討を行う．

2. オブジェクト検索のモデル

まず，本論文で取り上げる手法の前提となる，「オブジェクト検索」について，既存の研究を参照しつつ述べる．

2.1 オブジェクトのモデル

まず，オブジェクト検索を行うにあたって，オブジェクトのモデルに関して述べる．一般に，オブジェクト検索では，対象のオブジェクトのクラスのモデリングを先に行う．具体的には，対象となるクラスがどのような属性を持つかを規定する．これにより，オブジェクトを抽出したり集約したりする際に用いる，オブジェクト情報の構成要素を規定する．またその属性値について，先の問題点のところで述べた，値の特性などについても指定する．これらの情報は，同一オブジェクトかどうかを判定する際の尺度の基準となったり，属性値を適切に集約する際の判断の材料となる．

2.2 検 索 法

オブジェクト検索は，クエリとして単語群を入れたとき，その単語群が含まれているページを返す一般の Web ページ検索とは異なり，クエリとして，検索したいオブジェクトの属性値に関する指定を行う．例えば次のようなものが考えられる．

- （野球選手で）年齢が 30 歳以上，2007 年 6 月の所属が阪神タイガース

例えば野球選手というオブジェクトを扱うオブジェクト検索エンジンが存在したとき，そのオブジェクトは属性として名前，年齢，誕生日，身長といった人間一般の属性や所属球団，打率，タイトル数，入団年月日などの野球選手固有の属性を持つように定義されているはずである．このようなとき，各属性値（もっとも代表的なものは，名前）に関する指定を入力されることにより対象となるオブジェクトを決定し，それに関する集約された情報を返すというのが一般的なモデルである．このような複雑なクエリを処理することで，従来の Web ページ検索にない検索を行うことができる．

2.3 検索エンジンのアーキテクチャ

一方，実際の検索エンジンのアーキテクチャについてはいくつか考えられる．

文献 [2] や文献 [3] では，対象のオブジェクト群に関する情報は，事前にクロールしてオブジェクトの情報を収集，抽出，集約してデータベースとして蓄積しておき，その後ユーザからクエリを与えられたときは一般的なデータベースとしてふるまうというものである．

また，別の形式としては，ユーザがクエリを与えてから，検索エンジンなどを用いて対象に関する情報を探し出し，集約するアーキテクチャも考えられる．この場合，情報の集約の部分をユーザが操作できるので，よりユーザにあった集約を行うことができると考えられる [4].

いずれのアーキテクチャに於いても，このような検索を実現するためには，いくつかの代表的なコンポーネントが存在する．検索エンジンに必要なコンポーネントは次の通りである．

- オブジェクトの情報を含むページの収集
- オブジェクトの属性情報の抽出
- オブジェクトの属性情報の集約
- オブジェクトのランキング

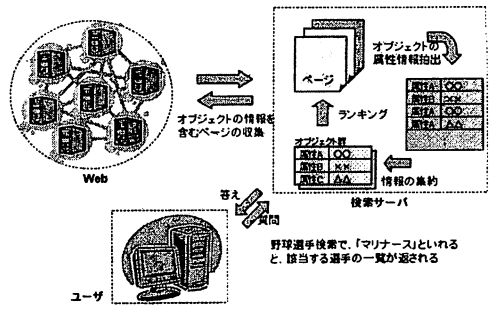

図 1 オブジェクト検索のモデル（クローラーを用いる場合）

まず，Web から該当のオブジェクトに関する情報が含まれているページを収集する．先に述べた 2 つのアーキテクチャのうち，前者の場合，オブジェクトの情報が含まれている可能性が高いページを効率よくクロールする必要がある．そのために，代表的なオブジェクトの例をいくつか与え，それに基づきオブジェクトの候補群を取得する．例えば，野球選手というクラスの場合では，野球選手には実際にどういう人物がいるかを推定する [2].

次に，収集されたページから，該当のオブジェクトとその属性に関する情報を抽出する．これは HTML や文章の書き方に依存したヒューリスティックなルールが用いられることが多い．ルールの多くはクラスに依存するので，対象とするクラスごとに作成しなければならない．

その後，抽出された属性情報群を適切に分類し，統合することで 1 つの整理された情報とする．別の情報源から得られる異なった情報は，本来が等しくなるべき場合でさえ，表記や内容が異なっていることが多いので，それらに対しうまく対応して重複した情報を集約する操作が必要となる．また，同姓同名などの，同一の名称を持つが異なる実世界のオブジェクトが存在する場合，抽出した属性情報を対応するオブジェクトごとに識別する必要がある．本論文で取り上げる手法はこのコンポーネントに含まれる．

最後に，各オブジェクトについてランク付けを行う [5]．クエリに該当するオブジェクトが複数あった場合，ランク付けを行う必要がある．

以上のコンポーネント同士の関わりを表したものが，図 1 である．

2.4 オブジェクトの属性情報集約の問題点

Web から抽出した情報を集約する際，同一の属性に対応する値であっても，様々な理由により異なった値を持っているものが存在し，集約が容易でないことが多い．以下にその原因を示す．

- 情報の誤り … 情報そのものが間違っており，そのため異なった値が得られた．
- 抽出法の誤り … プログラム的，アルゴリズム的原因により，誤った値が得られた．
- 表記のゆれ … 同じ値を指しているはずだが，表記が異

なっている．

- 別オブジェクトの参照 … 同じ名称を持つ別オブジェクトを参照してしまったため，異なった値が得られた．
- 複数の値を持つ属性値 … その項目が複数の値を持つことを許容するため，そのうち異なったものを得ることがあった．
- 定義のあいまいさ … その項目の値が厳密に定義されていない，または複数の説があるなどの理由により，解釈の違いによって異なるものとなっている．
- 情報の詳しさの違い … 細かく説明している情報と，大雑把に説明している情報がある．
- 時間の変化 … 属性値が時間によって変化しており，時間の異なる情報を参照したため，異なった値が得られた．

これらのうち情報の誤りと抽出法の誤りについて，文献[3]において誤りを低減する方式が提案されている．表記のゆれについては文献[6]など，数多くの研究がなされている．また，同名だが異なるオブジェクトの参照問題については，オブジェクト識別問題としてなどの様々な研究がある[7]．既存の研究では，別の名称を持つが同一のオブジェクトを指す問題や，同じ名称を持つが別のオブジェクトを指す問題については積極的に扱ってきた[8]が，オブジェクトの属性の時間による変化などはしないとして仮定しており，その問題について扱った論文は少ない．

3. 時間変化する属性のモデル

次に，このようなオブジェクトと属性を扱う上での問題点について述べる．時間によって属性値が変化するオブジェクトの中でも，属性値の数，属性値の有効期限，属性値が空の値を取りうるかどうか，など実際には様々なパターンが存在し，それらを分類して扱わなければ，単に時間順に並べて同じ情報を結合するだけでは誤った集約を行ってしまう可能性がある．例えばある野球選手の「タイトル」という属性値について，「1999年 ホームラン王」「2001年 ホームラン王」という情報を得たとき，これらを安易に繋げて「2000年 ホームラン王」といった補完をしてしまっては，誤った結論を導いてしまう．そこでまず，属性値の変化パターンについて述べる．

3.1 属性値の変化パターン

属性値の変化はあらゆるパターンで起こりうる可能性があるが，実際はごく一部の例外を除き，大多数が代表的パターンに沿った変化を行うと考えられる．以下にその代表的パターンについて示す．

3.1.1 値の変化しない属性値

オブジェクトが生成されると同時に決定され，以後，通常変更されない属性値．例えば人間でいうと血液型や生年月日に当たる．

3.1.2 時間変化するが値が必須な属性値

ある程度の間隔ごとに変化し，一つしか持たず，また常に一つ存在する属性値．例えば会社というオブジェクトで言えば，名前という属性値に当たる．

3.1.3 有効期限のある属性値

ある時点において発生し，その後一定期間を経て消滅する属性値．例えば企業というオブジェクトで言えば，商品というオブジェクトがそれにあたる．

3.1.4 蓄積型の属性値

ある時点において獲得され，以降消滅することはなく蓄積されてゆく属性値．例えば人物で言う著作物などにあたる．

3.1.5 時系列の属性値

主に数値について，非常に細かいタイミングで変化しており，1つの値として集約することが難しい属性値．例えば国というオブジェクトで言えば人口といったものが挙げられる．

3.2 時間情報のモデル

時間によって変化する可能性があるオブジェクト属性を考慮するため，各オブジェクトの属性に対し時間情報を付与する方法を考える．

一般に，Webから得られた属性に対する時間情報は，厳密には記述されていない．つまり例えば，「2006年」に「イチロー」の所属が「シアトルマリナーズ」であるという情報を得たとき，これらは必ずしも「2006年の間ずっと」シアトルマリナーズだったかどうかについては言及しないし（もちろんそうである確率が高いが），2005年や2007年にどうだったかについても完全な情報は得ることができない．この文字列からは得られる情報は，漠然とした推測のもとに成り立つ情報がある程度含まれているといえる．

このような，不明確な情報を不明確なまま処理し，あいまいさを表現するためには，時間情報を次のように不明確さを許したまま定義する必要がある．

- 開始年月日：日付 or 不明
- 終了年月日：日付 or 不明
- 期間：期間 or 不明

このようなオブジェクト情報を，いかにしてWebページから抽出し，作成するかについて，次の章で述べる．

4. 属性値の時間情報の抽出

この章では，Webから実際に，オブジェクトの情報する手法，特に属性値に対応する時間情報を取得する方法について述べる．例えば，「イチロー」が「新人王」という情報を得たとき，それがいつの情報であったかを，文章中の日付表現を解析することによって取得する．

手法としては主に，「情報の中に日付が明記されている場合，それを利用する」という点と，「情報の中に時間情報がない場合，暗黙的に文書が作成された時刻に有効な情報と仮定する」という2つが挙げられる．これらに必要な，主なコンポーネントについて挙げる．

- 文書からの日付表現の抽出
- 文書の書かれた時間情報の取得
- 文書内の日付の補完
- オブジェクトの属性情報に対応する日付情報の決定

4.1 文書中の日付表現の取得

これはマッチするパターンを書き連ねることで容易に抽出可能である．ただし，「○月×日」といった具体的に表現されているものだけでなく，「昨日」「今日」や「春ごろ」といったものまで含めると，非常に多くの数となる．「昨日」や「今日」はこれ単体だけでは情報として活用できないので，それを補完する必要がある．

4.2 文書の書かれた時間情報の取得

まず第一の目的は，文章の書かれた日付を取得することである．これは一般に，特に文章中に書かれている内容について明示的な指定がない限り，その内容は文章の書かれた日付の時点において有効であるからである．また，「昨日」や「今日」といった，補完の必要な表現が出現したとき，文書の書かれた日付が必要となることが多いためである．

まず，文書の書かれた日付情報を取得する，もっともシンプルな方法は，Web サーバと通信する際に得られる http ヘッダに存在する，Last-Modified の項目を取得することである．HTTP ヘッダは，Web 上での通信で最も一般的と思われる http プロトコルにおいて通信する際，クライアントとサーバの間で交わされるメタ情報である．そのうち，Last-Modified という項目に，Web ページの最終更新日時が含まれているので，それを取得すれば文書の作成日時を取得することができる．

しかし，これには問題点がいくつか存在する．1 つに，CGI や SSI などにより作成される Web ページは，この値を適切に返さないということである．これらの動的に生成されるページは，スクリプトの更新日時を返したり，スクリプトが作成したページの日付（つまり，クライアントがデータを要求した日付）を返すことがある．また，そうでなくとも，文書が後から少しだけ更新されるなどして，必ずしもファイルの日付が作成日時であるとは限らないという問題点もある．

そこで，別のアプローチも考えられる．直接的な方法として，文書内から日付表現を取得する方法がある．例えば，ヒューリスティックな方法として，「最終更新日」などと書かれた部分を取得する方法がある．しかしこれはニュースページなどから抽出するには有用であるが，カバーできないページが多い．別の方法としては，ページ内に存在する日付表現を取得するというものも考えられる．しかしこれは，未来について記述されている情報に対応できないなどの問題点がある．

以上とは別の問題点として，必ずしも文書全体が同一の日付において作成されているとは限らないということが考えられる．その最たる例が Blog である．Blog はセクションごとに明確に書かれた日付が異なり，「昨日」や「今日」の処理も区画別に行わなければならない．また，Blog は明確に各記事の日付が書かれているため，適切に処理すれば解決できるはずである．

4.3 文書中の日付表現の補完

文書の作成された日付が解決した場合，次に文書中に存在する「昨日」や「今日」を解決することにより，文書中から多くの日付表現を得ることができるようになる．このような手法は文献 [4] などで提案されている．

4.4 情報への日付情報の付与

次に，これらの作業により文書中に存在する日付表現が多く発見されているはずである．そこで，文書中の情報と日付表現を関連付ける方法を以下で考える．関連付けるパターンには次の 2 通りが考えられる．

- 情報そのものに，日付表現が含まれている 例：2001 年，イチローはマリナーズに入団した．
- 日付表現がなく，事実のみが記されている 例：Blog 記事の見出しイチローがマリナーズ入団

前者の場合，日付と関連付けるのは非常に容易である．文章中に含まれている日付表現をその情報の日付とすれば良い．しかし，一般に「同じ文内」に日付表現が入っているものは少なく，その文の前の文や同じ段落内，あるいは日付表現を補完してようやくわかるものが存在する．これらを利用することで，このようにして得られる情報を増やすことができる．

後者に対しては，その文書の日付を以って情報の有効時間とするのが妥当であると考えられる．

4.5 日付情報の展開

以上により，「属性情報」と「時間表現」のペアが得られた．最後にこの時間表現を，先ほど述べた形式に変換することを考える．変換する例について述べる．

例えば「1999 年から」という表現があった場合，該当する期間の開始時間は 1999 年である．期間や終了時間はわからないので，不明としておく．「1999-2005」という表現があった場合，開始時間が 1999 年，終了時間が 2005 年である．期間は 6 年間である．

後に，属性値に起因する特性によりさらに時間情報が付加される場合がある．その後，同一と思われる属性値同士の時間範囲を結合してゆき，一つの属性値の有効期間としてまとめる．

5. 属性値が時間変化するオブジェクト情報の集約

次に，以上の手法によって作成された時間付きの情報に対して，どのような処理を行うかについて述べる．識別や集約において，3.1 章で分類した属性の時間情報を用いることで，時間によって変化する属性値に関連した問題点を解決する．

5.1 属性の時間的特性の付加

日付表現からは属性値の有効期間について一定の情報が得られるが，属性値に固有の知識を使うことでも，暗黙的な情報を得られることがある．ここではそのような，属性そのものに起因する暗黙的な時間的情報を補完する方法について述べる．

オブジェクト検索の対象となるオブジェクトの属性には，さまざまな種類が存在するが，それについて 3.1 章で分類を行った．ここでは，各属性がそれらのどれにあてはまるかが 3.2 章により情報を得られているものとする．この場合，属性の時間的変化の特性の種類によっては，4. 章で得られた情報に対しさらに情報を付与することができる．

具体例を示す．2006 年 8 月の時点で，ある企業のオブジェクトについて，「設立年月日」という属性の属性値が「○○年×

月」であったとする．すると，この属性値は，2006年8月の時点のみならず，通常あらゆる時点において有効な事実であるという風に解釈することができる．

5.2 属性値の時間変化に基づくオブジェクトの識別

ここまでで得られた情報について，これらを分類し，同一オブジェクトのもの同士でまとめる方法について述べる．この時点で集められた情報について，まだ意図しないものが含まれている場合が考えられる．以下にその例を挙げる．

- 同姓同名の別人物の情報が含まれている．
- 抽出部分などのアルゴリズム的誤りから，本来関係のない誤った情報が含まれている．
- Webソースの内容の誤りから，似通っているが矛盾のある情報が入っている．

以上を処理するために，次のような手法を考える．オブジェクトのモデリングの際に決定した情報を用いて，各情報同士に対し，同一人物かどうかの自然さの距離尺度を定義する．具体的には，次のように定義する．

- 同一人物であるとの認定を一定の確度で行えそうなもの…類似度 高い
- 同一人物かどうか，判定する材料が無いもの…類似度 低い
- 矛盾した内容を持つもの…類似度 0

これらを元にクラスタリングを行う．オブジェクト情報の抽出法にも依存するが，同じページに存在する情報が同じ人物であるという情報を加えれば（この仮定は厳密には必ずしも常に真ではないが），うまく情報をまとめることができる．その結果，同一人物同士でまとまったクラスタを作成することができる．(図を図??に示す.)

また，属性値の時間変化を利用することでも，同一オブジェクトかどうかの識別を行うことができる．図2に例を示す．あるオブジェクト4つが存在し，それらの属性値がそれぞれある期間においてさまざまな値をとっていたとする．もしその属性の時間的変化の特性が，複数の値を取らずに，また空白値を取らない属性であれば，これら4つのオブジェクトが同一オブジェクトかどうかは，その属性値が取る期間の被覆具合で判定できる．重複する部分があれば明らかにおかしいし，重複せず，かつ空白期間を作らずに結合できれば，それらが同一オブジェクトと考えるのは非常に自然である．

5.3 時間変化する属性値の集約

以上を元に，同一人物と思われる情報群を得ることができた．最後に，情報の集約と提示を行う．

主に行われる処理は，時間変化していると思われる情報同士の結合と，矛盾した情報の処理，それに抽象度の異なるもの同士の処理である．

具体的に例を挙げながら示す．

- 1998年 オリックス・ブルーウェーブ
- 1999年から マリナーズ

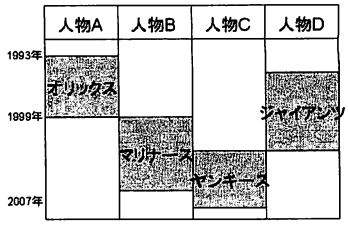

図2 属性値の時間変化による識別

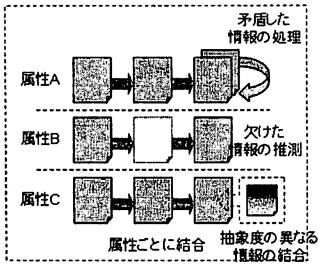

図3 情報の結合モデル図

- 2003年 マリナーズ

ある野球選手の，所属球団という属性について，上のような情報が抽出されたと仮定する．この場合，次のような推論を行う．まず，野球選手というオブジェクトのモデリングの際，所属球団という属性についての有効期間の平均と分散に関する情報を与えておく．この場合に即して言えば，同一球団にとどまる長さの平均と分散である．それと，属性値の数（この場合で言えば，所属球団の数）の平均と分散も必要である．これらの情報を用いれば，まずこの野球選手の所属球団という属性が，オリックスとマリナーズの2つが同時に存在した可能性が否定できる．また，1998年から1999年の間にオリックスとマリナーズ以外の球団に所属した可能性は低いと結論できる．最後に，1999年から2003年の間に，マリナーズ以外の球団に所属した可能性は低い．以上より，次のようにまとめあげることができる．

- 不明-1998年 オリックス・ブルーウェーブ
- 1999年-2003年以降 マリナーズ

このような作業によって，情報を統合する．

5.4 時間変化する属性値の提示

最後に，以上より得られたものをまとめた最終的な出力のイメージ図を示す．

5.4.1 複数のオブジェクトの出力

検索により得られた結果は，同名で別のオブジェクトが検索されるなどの理由により複数のオブジェクトが得られている可能性がある．その場合，オブジェクトごとに結果を提示する(図4)

5.4.2 時間によって変化する属性の出力

時間によって変化する属性の出力例を示す．時間順に沿って属性値を出力する．

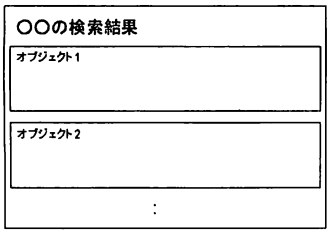

図4 出力例1

名称	○○株式会社	
社長	1992.4 - 1996.3	○田×夫
	1996.4 - 2000.3	△下□一
	2000.4 -	○○株式会社
	:	

図5 出力例2

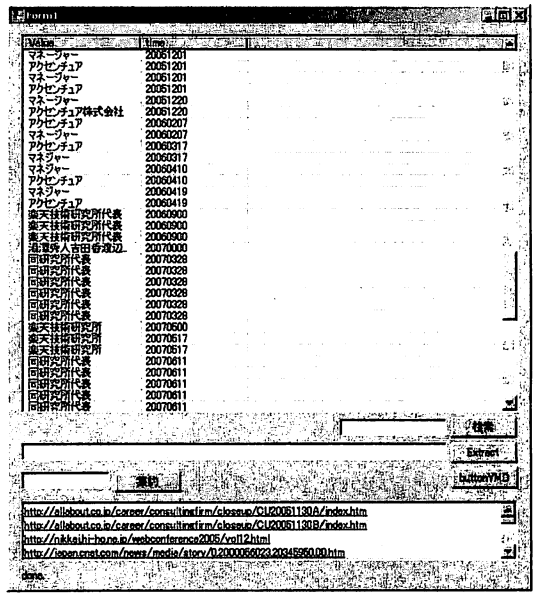

図6 属性抽出プログラムプロトタイプ

6. プロトタイプ

現在の段階で作成されているプロトタイプシステムを図6に示す．このシステムは人物を対象オブジェクトのクラスとし，ユーザが入力したクエリ，すなわち人名を用いてWeb検索を行い，該当する人物の「所属」に該当する属性をヒューリスティックなルールを用いて抽出する．今回はヒューリスティックとして，形態素解析を用いて，人名の直前，直後に現れる名詞を抽出している．所属を表す文字列の近傍の日付表現を取得し，木村ら[4]の提案した，日付表現の補完を行い，属性値の日付としている．今後，5章で提案した，属性のタイプ（例えば職業の場合，時間的に変化するが値が必須である）ごとに情報を集約する仕組みを実装予定である．

7. まとめと課題

本論文では，オブジェクト検索におけるモデルの中で，時間変化する属性値について取り上げ，それらを適切に扱うことでオブジェクトの識別や検索を正しく行うための手法について考察した．

残った課題としては，一つに不明確な情報に対する取り扱いが挙げられる．Webから抽出された情報は不完全なものが多く，データを比較したり統合したりする際にもそれを考慮しなければいけないことが多い．また，空白のデータを推論して補完する際にも，確実ではない結論が出るため，それを扱う必要がある．推論の程度を定量化するためにも，不明瞭さを扱う必要がある．不明瞭さを定量的に数値化したりするなど工夫が必要である．

また，上の話題と関連し，Webから取得した情報は誤りがあるものも少なからず存在し，情報を集約する上での障害となる．よって，その信頼性を評価することが欠かせない作業となる．また，システムが推論して得た情報も，誤りを含む可能性を持つ．ユーザにより正確な情報を伝え，誤解を与えないためにも，情報の信頼性を評価し，ユーザに提示する機能が必要であると考えられる．

謝辞

本研究の一部は，科学研究費補助金（課題番号19700091,18049041,18049073）および文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」（研究代表者：田中克己）によるものです．ここに記して謝意を表します．

文　献

[1] R. Guha and A. Garg. *Disambiguating people in search.* Stanford University, 2004.

[2] Zaiqing Nie, Ji-Rong Wen, and Wei-Ying Ma. Object-level vertical search. In *CIDR2007*, 2007.

[3] Zaiqing Nie, Yunxiao Ma, Shuming Shi, Ji-Rong Wen, and Wei-Ying Ma. Web object retrieval. In *WWW2007*, 2007.

[4] Rui Kimura, Satoshi Oyama, Hiroyuki Toda, and Katsumi Tanaka. Creating personal histories from the web using namesake disambiguation and event extraction. In *ICWE2007*, 2007.

[5] Zaiqing Nie, Yuanzhi Zhang, Ji-Rong Wen, and Wei-Ying Ma. Object-level ranking: Bringing order to web objects. In *WWW2005*, 2005.

[6] Mikhail Bilenko and Raymond J.Mooney. Adaptive duplicate detection using learnable string similarity measures. In *SIGKDD2003*, 2003.

[7] 白砂健一, 小山聡, 田島敬史, 田中克己. Webの構造情報とプロファイル抽出を用いたオブジェクト識別. In *DEWS2006*, 2006.

[8] 外間智子, 北川博之. Webデータを用いた人物の呼称抽出. In *DBSJ Letters 2006*, 2006.