1B1-4

属性値が時間変化するオブジェクトを識別する確率モデル

Probabilistic Models for Identifying Objects with Time-Varying Attribute Values

小山 聡 Satoshi Oyama　白砂 健一 Kenichi Shirasuna　田中 克己 Katsumi Tanaka

京都大学大学院 情報学研究科 社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

We have developed a method for determining whether data found on the Web are for the same or different objects that takes into account the possibility of changes in their attribute values over time. By giving a specific form to the distributions of time-varying attributes, we can calculate the similarity between given data and identify objects by using agglomerative clustering on the basis of the similarity. Experiments showed that the proposed method improves the F-measure of object identification.

1. はじめに

実世界のオブジェクトの多くは，時間とともに属性値が変化する．人物であれば，数年経てば所属や連絡先が変わることは珍しくないし，企業でも，所在地や代表者の名前はしばしば変化する．管理されたデータベースにおいては，属性の情報が更新されたり，古い情報が削除されたりして整合性が保たれるが，Web など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．

現在，オブジェクト検索エンジン [Nie 07] と呼ばれる，あるクラスのオブジェクトに関する情報を，ページ単位ではなくオブジェクト単位で検索できる新しいタイプの検索エンジンの研究が行われている．多くのオブジェクト検索エンジンでは，対象とするクラスのオブジェクトが持つ属性やオブジェクト間の関連をスキーマとして定義し，スキーマに基づいて Web から情報を抽出，集約することが行われる．

オブジェクト検索を実現するためには，Web ページから抽出した各データを，実際のオブジェクトに対応させるオブジェクト識別と呼ばれる処理が必要になる．オブジェクト識別は，基本的にはデータの属性値から類似度を計算し，クラスタリングすることで実現される．これは Web 上の情報に限らず，従来からデータベースにおいて重複除去などの目的で用いられてきた技術である．しかし，Web 上の情報に対してオブジェクト識別を行うことはより困難である．その理由の一つは，先に述べたように，オブジェクトの情報が変化することがある．従来のオブジェクト識別の研究でしばしば対象オブジェクトとされた学術論文などでは，オブジェクトの属性値（著者名や掲載雑誌名）はオブジェクトが生成された時点で確定し，時間的に変化しない．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，代表者や資本金額など）が時間に従って頻繁に変化することが多い．Web から収集された情報には，属性値が時間変化した同一オブジェクトの異なる時点での情報が混在し，集約の際に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．

一方，人物や企業の場合，現在の情報だけでなく，過去の情報も網羅的に Web から収集したい場合がある．オブジェクト検索の今後の重要な方向として，オブジェクト履歴検索 [Kimura 07] が考えられる．今後 Web 上や Web アーカイブ中に多くの情報が蓄積されていくに従い，このような情報要求は増加し，時間変化するオブジェクトを識別する技術は重要になっていくと考えられる．

本稿では，時間変化するオブジェクトを精度よく識別するための手法を検討する．具体的には，観察されたデータが，別のオブジェクトに由来するものか，時間変化した同一オブジェクトのものかを表す確率モデルを構築し，このモデルに基づいてオブジェクト識別を行う．

連絡先: 〒 606-8501 京都市左京区吉田本町 京都大学大学院 情報学研究科 社会情報学専攻 小山 聡

2. 基本モデル

オブジェクトに関する 1 回の観測を $(\boldsymbol{x}_i, t_i)$ で表す．ここで，

$$\boldsymbol{x}_i = (x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(N)})$$

である．各 $x_i^{(n)}$ はオブジェクトの観測された属性値，N は異なる属性の数である．t_i はこのデータが観測された時刻を表す．また，観測 $(\boldsymbol{x}_i, t_i)$ を生成したオブジェクトを一意に識別する識別子（一般には値が未知である）を o_i で表す．2 つの観測 $(\boldsymbol{x}_i, t_i)$ と $(\boldsymbol{x}_j, t_j)$ が同じオブジェクトから生成されたとき，$o_i = o_j$ であり，別オブジェクトから生成されたとき $o_i \neq o_j$ である．

2 つの観測 $(\boldsymbol{x}_i, t_i)$ と $(\boldsymbol{x}_j, t_j)$ が与えられた時，それが同一オブジェクトから生成されたものである確率 $p(o_i = o_j|(\boldsymbol{x}_i, t_i), (\boldsymbol{x}_j, t_j))$ と別のオブジェクトから生成されたものである確率 $p(o_i \neq o_j|(\boldsymbol{x}_i, t_i), (\boldsymbol{x}_j, t_j))$ を比較したい．そのために，オッズ比の対数

$$\log \frac{p(o_i = o_j|(\boldsymbol{x}_i, t_i), (\boldsymbol{x}_j, t_j))}{p(o_i \neq o_j|(\boldsymbol{x}_i, t_i), (\boldsymbol{x}_j, t_j))} \tag{1}$$

を用いる．これは，同じオブジェクトである可能性が高いとき，大きな正の値をとり，異なるオブジェクトである可能性が高いとき，大きな負の値をとる．

ベイズの定理を用いると，式 (1) は

$$\log \frac{p((\boldsymbol{x}_i, t_i), (\boldsymbol{x}_j, t_j)|o_i = o_j)}{p((\boldsymbol{x}_i, t_i), (\boldsymbol{x}_j, t_j)|o_i \neq o_j)} + \log \frac{p(o_i = o_j)}{p(o_i \neq o_j)}$$

と書ける．第 1 項は同一オブジェクトの場合と異なるオブジェクトの場合の対数尤度比であり，時間を考慮しない場合は Fellegi

と Sunter[Fellegi 69] による従来のレコード同定の定式化と等価である．第 2 項は 2 つの観測が同一オブジェクトから生成される事前分布と別オブジェクトから生成される事前分布の対数比であるが，これは観測データに依存しない．そのため，この項は式 (1) を観測の間の類似度としてクラスタリングなどで用いる場合には無視することができる．

ここで計算を簡単にするために，オブジェクトが同一か否かが与えられた時の属性間の独立性を仮定する．すると，観測間の類似度 similarity$((\boldsymbol{x}_i, t_i), (\boldsymbol{x}_j, t_j))$ は，各属性 n に対して個別に類似度

$$\begin{aligned}&\text{similarity}((x_i^{(n)}, t_i), (x_j^{(n)}, t_j))\\ &= \log \frac{p((x_i^{(n)}, t_i), (x_j^{(n)}, t_j) | o_i = o_j)}{p((x_i^{(n)}, t_i), (x_j^{(n)}, t_j) | o_i \neq o_j)} \end{aligned} \tag{2}$$

の値を計算し，その和を取ることで求めることができる．

さて，各属性で式 (2) の値を求める方法を考えよう．以降では記述の簡単化のために，属性を示す添え字 (n) を省略して議論する．まず，分母について考える．これは，i と j が別のオブジェクトであるときに，i の属性値が時刻 t_i で x_i，j の属性値が時刻 t_j で x_j をとる確率である．ここで，オブジェクトの母集団において，属性値の分布 $p((x,t))$ は時間的に不変である，すなわち，$p((x,t)) = p(x)$ であると仮定する．また，別オブジェクトのときに観測される属性値が条件付き独立であるとすると，式 (2) の分母は

$$p(x_i)p(x_j)$$

と近似できる．これは，オブジェクトの母集団から 2 つオブジェクトをサンプリングしたときに，それらの属性値が x_i と x_j となる確率である．

次に，式 (2) の分子について考える．ここで，$t_i < t_j$ であるとする．すると，あるオブジェクトの属性が時刻 t_i で値 x_i を取り，その後時刻 t_j までに値が x_j に変化したと解釈することができる．ここでもオブジェクトの母集団での属性値の分布は変化しないとすると，時刻 t_i で属性値 x_i が観測される確率は t_i に依存せずに単に $p(x_i)$ となる．また，時間が $t_j - t_i$ 経過した後の x_j の分布を

$$q((x_j, t_j)|(x_i, t_i))$$

とする．これらを用いて式 (2) の分子は

$$p(x_i)q((x_j, t_j)|(x_i, t_i))$$

と書ける．これを用いると，結局

$$\begin{aligned}&\text{similarity}((x_i, t_i), (x_j, t_j)) = \log \frac{p((x_i, t_i), (x_j, t_j) | o_i = o_j)}{p((x_i, t_i), (x_j, t_j) | o_i \neq o_j)}\\ &= \log \frac{p(x_i)q((x_j, t_j)|(x_i, t_i))}{p(x_i)p(x_j)} = \log \frac{q((x_j, t_j)|(x_i, t_i))}{p(x_j)}\end{aligned}$$

となる．

3. 類似度の計算

ここで，観測間の類似度を計算するために，p と q の関数形を具体的定めることを考える．これらのモデルは，オブジェクトのクラスごとにドメインの知識を用いて事前に設計する必要があるが，以下では，カテゴリ属性と数値属性についての簡単なモデルを例として与える．

3.1 カテゴリ属性

スポーツ選手の所属チームなどは，時間が経過するにつれて値が変化するカテゴリ属性の例である．一般には，各属性値間の遷移確率を指定する必要がある．ここでは，簡単な場合として，属性値のとる確率および属性間の遷移確率が一様な場合を考えよう．たとえば，スポーツ選手の例で，L 個の球団があり，各球団の所属選手数は均等だとしよう．また，単位時間あたりに，チームを移動する割合を r としよう．すると，

$$\begin{aligned}&q((x_j, t_j)|(x_i, t_i))\\ &= \begin{cases} (1-r)^{t_j - t_i} + \frac{1}{L}\left(1 - (1-r)^{t_j - t_i}\right) & (x_i = x_j) \\ \frac{1}{L}\left(1 - (1-r)^{t_j - t_i}\right) & (x_i \neq x_j) \end{cases}\end{aligned}$$

という分布が得られる．

$p(x) = \frac{1}{L}$ であるので，いずれの場合も $t_j - t_i \to \infty$ で

$$\text{similarity}((x_i, t_i), (x_j, t_j)) = \log \frac{q((x_j, t_j)|(x_i, t_i))}{p(x)} \to 0$$

となる．これは，十分時間が経過した後では，属性の情報が意味を失うことを表している．

3.2 数値属性

次に，時間的に変化する数値属性について考える．たとえば年収などは，平均的には毎年一定の割合で増加していっても，個々の例では減少したり，平均より大きく増加したりといったばらつきが観測される．簡単な例として，時間 $t_j - t_i$ が経過した後，属性値 x_j の分布が，平均 $\mu_{x_i, t_j - t_i}$，分散 $\rho^2_{t_j - t_i}$ の正規分布にしたがう場合を考える．ここでは，平均的には時間に比例して属性値の値が増減すると考え，$\mu_{x_i, t_j - t_i} = x_i + \alpha(t_j - t_i)$ とおく．α は単位時間当たりの変化率である．

時刻 $t_j - t_i = 0$ のとき，属性値 x_j の値は x_i に一致する必要がある．一方，十分時間が経過した $t_j - t_i \to \infty$ では，x_j の分散はある一定値 ρ^2 に収束するようにしたい．そのために，ここでは

$$\rho^2_{t_j - t_i} = \frac{\beta(t_j - t_i)}{1 + \beta(t_j - t_i)}\rho^2$$

とおく．$\beta > 0$ は単位時間当たりの変化率である．これらから，以下のような正規分布が得られる．

$$q((x_j, t_j)|(x_i, t_i)) = N\left(x_i + \alpha(t_j - t_i), \frac{\beta(t_j - t_i)}{1 + \beta(t_j - t_i)}\rho^2\right)$$

この式を見ると，$t_j - t_i = 0$ で，$x_j \neq x_i$ の場合，分散が 0 になることが分かる．これは，同一時点で同じオブジェクトに異なる値が観測されることがあり得ないことを表している．しかし数値属性の場合は，たとえ同一時点であっても観測誤差によって異なる値が得られる場合がある．そうすると，本来同一オブジェクトと判定されるべきものが異なるオブジェクトと判定されることになる．そのため，ある程度の測定誤差を許容する必要がある．誤差 δ の分布が $N(0, \sigma^2_{\text{error}})$ に従うとする．時刻 t_j での観測値は，真の値と観測誤差の和であり，これらが独立であるとすると，

$$\begin{aligned}&q((x_j, t_j)|(x_i, t_i))\\ &= N\left(x_i + \alpha(t_j - t_i), \frac{\beta(t_j - t_i)}{1 + \beta(t_j - t_i)}\rho^2 + \sigma^2_{\text{error}}\right)\end{aligned} \tag{3}$$

となる．

企業の規模（資本金額）のようなある種の属性は，正規分布から離れた，歪んだ分布を持つ．また，時間変化においても，加法的ではなく，乗法的な変化をすることが多い．すなわち，時刻 t_j での値が $x_j \sim x_i\beta^{(t_j-t_i)}$ 典型的に従う．この場合，変数の対数を取ることで，$\log x_j \sim \log x_i + (t_j - t_i)\log\beta$ と線形な式に置き換えることができる．実際，企業規模などの多くの値が対数正規分布に従うことが知られている．このような属性を扱う場合には，我々は属性値の対数をとることで，式 (3) を用いた問題に帰着することができる．

3.3 属性値誤りの扱い

Web においてはページ自体における記述の誤りや情報抽出の誤りによって，本来の値とは異なった属性値が観測データに混入する可能性がある．同一オブジェクトの同一時点のカテゴリ属性や時間変化しない属性が異なる値を取らないという制約を厳密に満たそうとすると，誤った属性値によりオブジェクト識別の精度が大きく低下する可能性がある．ここで，情報の誤り率を γ とすると，本来同一オブジェクトに由来する 2 つのデータであるのに，どちらに誤りが生じる確率は，誤りが 2 重に起きて結果的に同じ値を取る確率は無視すると，およそ 2γ となる．また，誤りが生じた際に，属性値 x_j が選択される確率が母集団での属性値の分布に比例すると考えると，誤って属性値 x_j が観測される確率は $2\gamma p(x_j)$ で表される．誤りが生じない場合は $q((x_j,t_j)|(x_i,t_i))$ に従って属性値が選択されるので，誤りを考慮した場合の分布は

$$q'((x_j,t_j)|(x_i,t_i)) = (1-\gamma)q((x_j,t_j)|(x_i,t_i)) + 2\gamma p(x_j) \quad (4)$$

とすれば良い．実際，誤りによって $q((x_j,t_j)|(x_i,t_i))$ が非常に小さい，ほとんどあり得ないような値が得られたとしても，$\mathrm{similarity}((x_i,t_i),(x_j,t_j)) = \log 2\gamma$ となり，類似度が $-\infty$ になってしまう問題を避けることができる．

4. クラスタリング

観測の集合をクラスタリングすることで，オブジェクト識別を行う．ここでは階層的クラスタリングのアプローチを採用する．2 つの観測の間の類似度を，2 つのクラスタ（観測の集合）の間の類似度に拡張する．類似度は，異なるクラスタに属する観測が同一オブジェクトのものである（クラスタをマージした後の）尤度と，異なるオブジェクトのものである（クラスタをマージする前の）尤度の対数比である．

2 つのクラスタ $C' = \{(\boldsymbol{x}'_1,t'_1),(\boldsymbol{x}'_2,t'_2),\ldots,(\boldsymbol{x}'_K,t'_K)\}$ と $C'' = \{(\boldsymbol{x}''_1,t''_1),(\boldsymbol{x}''_2,t''_2),\ldots,(\boldsymbol{x}''_L,t''_L)\}$ が与えられたとする．ここで，$t'_1 \le t'_2 \le \cdots \le t'_K$ かつ $t''_1 \le t''_2 \le \cdots \le t''_L$ とする．次に，この 2 つのクラスタをマージし，新しい 1 つのクラスタ $C = C' \cup C'' = \{(\boldsymbol{x}_1,t_1),(\boldsymbol{x}_2,t_2),\ldots,(\boldsymbol{x}_{K+L},t_{K+L})\}$ にすることを考える．ここで，$t_1 \le t_2 \le \cdots \le t_{K+L}$ とする．マージ前とマージ後の尤度比は

$$\begin{aligned}
&\mathrm{similarity}(C',C'') \\
&= \log p(\boldsymbol{x}_1) + \sum_{m=2}^{K+L} \log q((\boldsymbol{x}_m,t_m)|(\boldsymbol{x}_{m-1},t_{m-1})) \\
&\quad - \left\{ \log p(\boldsymbol{x}'_1) + \sum_{k=2}^{K} \log q((\boldsymbol{x}'_k,t'_k)|(\boldsymbol{x}'_{k-1},t'_{k-1})) \right\} \\
&\quad - \left\{ \log p(\boldsymbol{x}''_1) + \sum_{l=2}^{L} \log q((\boldsymbol{x}''_l,t''_l)|(\boldsymbol{x}''_{l-1},t''_{l-1})) \right\}
\end{aligned}$$

で計算できる．上の式は，

$$\begin{aligned}
&\mathrm{similarity}(C',C'') \\
&= \sum_{m=2}^{K+L} \mathrm{similarity}\left((\boldsymbol{x}_{m-1},t_{m-1}),(\boldsymbol{x}_m,t_m)\right) \\
&\quad - \sum_{k=2}^{K} \mathrm{similarity}\left((\boldsymbol{x}'_{k-1},t'_{k-1}),(\boldsymbol{x}'_k,t'_k)\right) \\
&\quad - \sum_{l=2}^{L} \mathrm{similarity}\left((\boldsymbol{x}''_{l-1},t''_{l-1}),(\boldsymbol{x}''_l,t''_l)\right)
\end{aligned}$$

と同一であることを示すことができる．すなわち，クラスタ内で時間的に隣接する観測間の類似度の和から計算することができる．

最も類似度の高いクラスタの対をマージすることを，類似度が与えられた閾値を下回らない限り繰り返すことで，オブジェクト識別を行う．

5. 実験

以上の手法の効果を検証するために，時間情報を考慮した場合のオブジェクト識別の精度と，考慮しない場合のオブジェクト識別の精度を比較する実験を行った．本稿では，プロスポーツ選手および企業を対象クラスとした．これらのクラスに対して，属性値の母集団での分布と時間変化のパターンを指定したスキーマを作成した．表 1 にプロスポーツ選手のスキーマを例として示す．オブジェクト識別が困難となるのは，オブジェクト名が曖昧な場合が多い．そこで，表 2 に示すような，同姓同名の選手が存在する人名と類似な名前の企業が存在する企業名を対象オブジェクト名として選択した．

オブジェクト検索エンジンにおいては，オブジェクト識別は情報抽出を行った結果に対して行われるため，情報抽出の精度のオブジェクト識別の精度への影響を調べる必要がある．また，ページの時間分布も収集に用いた検索エンジンや Web アーカイブによって大きく異なる．そこでまず，人手でオブジェクトの履歴を年表として復元し，そこから観測データをサンプリングすることでテストデータを作成した．

具体的には，表 2 の名前を含むページを，検索エンジン（Google）および Web アーカイブ（Wayback Machine）から収集した．プロスポーツ選手については英語のページを，企業については日本語のページを用いた．人手でオブジェクトの同一性を判定し，属性情報および時間情報を抽出することで，年表を作成した．

次に，この年表から以下の手順でテストデータを抽出した．

1. まず，オブジェクトをサンプリングする．実際にはオブジェクトの観測頻度自体も大きく異なる．ここでは，Web での出現頻度に基づいて選択した．

2. 次に，観測時点をサンプリングする．ここで，2 つの異なるシナリオを考慮した．「検索エンジン」シナリオでは，は観測時点は指数分布（ここでは指数 0.4）に従い，最近の時点が多く観測されるようにした．「Web アーカイブ」シナリオでは，観測時点は一様にサンプリングした．

3. その後，観測される属性をサンプリングする．属性によって，観測される確率は異なる．ここでも，Web での観測頻度に基づいて確率を定めた．

表 1: プロスポーツ選手のスキーマ

属性	カテゴリ/数値	時間変化	母集団での分布
身長	数値 (σ_{error} =3)	定数	$p(x) = N(72, 10^2)$
体重	数値 (σ_{error} =3)	線形 ($\alpha = 0, \beta = 2 \times 10^{-3}, \rho^2 = 10$)	$p(x) = N(200, 15^2)$
生年月日	カテゴリ	定数	$p(x) = 10^{-5}$
投	カテゴリ	定数	$p(\text{right}) = 0.8, p(\text{left}) = 0.2$
打	カテゴリ	定数	$p(\text{right}) = 0.8, p(\text{left}) = 0.2$
出身地	カテゴリ	定数	$p(x) = 10^{-5}$
出身高校	カテゴリ	定数	$p(x) = 10^{-5}$
出身大学	カテゴリ	定数	$p(x) = 10^{-5}$
年齢	数値 ($\sigma_{\text{error}} = 1$)	線形 ($\alpha = 1, \beta = 0$)	$p(x) = N(25, 5^2)$
チーム	カテゴリ	ランダム ($r = 0.1$)	$p(x) = 1/30$
経験	数値 ($\sigma_{\text{error}} = 1$)	線形 ($\alpha = 1, \beta = 0$)	$p(x) = N(3, 4^2)$
氏名	カテゴリ	定数	$p(x) = 1/5$
デビュー年	カテゴリ	定数	$p(x) = 1/5$
ポジション	カテゴリ	ランダム ($r = 0.01$)	$p(x) = 1/10$

表 2: F の最大値 (一様な時間分布)

オブジェクト名	時間変化を考慮しない	時間変化を考慮
Mark Johnson	0.8529	0.9983
Mike Johnson	0.5169	0.9737
Matt Smith	0.8015	0.9862
Steve Smith	0.7016	0.9923
James Williams	0.6915	0.9539
平均	0.7129	0.9809
日立	0.9389	0.9683
JR	0.7297	0.9174
三井住友	0.8479	0.9089
NTT	0.8789	0.9076
東京三菱	0.9203	0.9424
平均	0.8631	0.9289

4. 最後に，情報抽出の誤りをデータに反映するために，一定の割合で実際の値を例外値で置き換えた．データの誤り率のデフォルト値は $\epsilon_{\text{data}} = 0.01$ とした．

以上の手順で，各オブジェクト名に対して 100 個の観測からなる 5 つの独立なテストセットを作成した．ベースラインとして，時間を考慮しないスキーマ，すなわち，r，α，β を 0 と置いたものを用い，最短距離法でクラスタリングを行った結果と比較した．ベースラインにおいても，式 (4) を用いて情報抽出誤りへの対応は行った．

類似度の閾値を変化させることで，クラスタ数を変化させ，各クラスタ数での再現率と適合率，F 値を計算した．再現率-適合率曲線上での F の最大値を表 2 に示す．観測時点の分布は精度に大きな違いを与えなかったので，ここでは一様分布の場合だけを示している．手法間で差が顕著でない例もいくつかあるが，多くの例において，属性の時間変化を考慮した場合が考慮しない場合よりも高い F 値を達成している．表 3 にデータの誤り率を 10 倍にした場合の結果を示す．表 2 と比較すると，データの誤り率が大きくなるとクラスタリングの F 値は小さくなることがわかる．しかし，類似度計算で用いる誤り率の見積もり ϵ を大きくすることで，ある程度精度を維持できることが分かる．

6. おわりに

人物や組織などの実世界のオブジェクトは，時間とともにその属性値が変化する．Web においては，過去の古い情報と新しい情報が共存しているため，精度の良いオブジェクト識別を実現するためには，属性値の時間変化を考慮するする必要がある．本論文では，観測された異なる属性情報が，異なるオブジェクトに由来するものなのか，同じオブジェクトの属性が変化したものかを区別する枠組みを提案した．また，属性のタイプごとに，属性値の時間変化のパターンを表す確率モデルの例

表 3: $\epsilon_{\text{data}} = 0.1$ のテストセットに対する F の最大値（時間変化を考慮，一様な時間分布)

オブジェクト名	誤り率の見積もり ϵ			
	0.1	0.01	0.001	0
Mark Johnson	0.9735	0.9587	0.9553	0.8885
Mike Johnson	0.8911	0.8752	0.8565	0.8146
Matt Smith	0.9359	0.9117	0.9021	0.8309
Steve Smith	0.9110	0.9160	0.9163	0.8790
James Williams	0.7455	0.7369	0.7098	0.6398
平均	0.8914	0.8797	0.8680	0.8106
日立	0.9184	0.9245	0.9218	0.9143
JR	0.8552	0.8214	0.8096	0.7132
三井住友	0.8152	0.8229	0.8416	0.7754
NTT	0.7516	0.7543	0.7534	0.7402
東京三菱	0.8716	0.8729	0.8784	0.8076
平均	0.8424	0.8392	0.8410	0.7901

を与えた．この確率モデルから計算される観測間の類似度を用いて，時間変化するオブジェクトの識別を行う階層的クラスタリングアルゴリズムを示した．現在は，属性値の分布をドメインの知識を用いて事前に与える必要があるが，今後は Web から得られるデータを用いて，自動的に分布を獲得する研究を行う予定である．

謝辞

本研究の一部は，科学研究費補助金（課題番号 18049041, 19700091），文部科学省「知的資産のための技術基盤」プロジェクト，京都大学 GCOE プログラム「知識循環社会のための情報学教育研究拠点」，およびマイクロソフト産学連携研究機構 CORE 連携研究プロジェクトによる．

参考文献

[Fellegi 69] Fellegi, I. P. and Sunter, A. B.: A Theory for Record Linkage, *Journal of the American Statistical Association*, Vol. 64, No. 328, pp. 1183–1210 (1969)

[Kimura 07] Kimura, R., Oyama, S., Toda, H., and Tanaka, K.: Creating Personal Histories from the Web using Namesake Disambiguation and Event Extraction, in *Proc. ICWE 2007*, pp. 400–414 (2007)

[Nie 07] Nie, Z., Ma, Y., Shi, S., Wen, J.-R., and Ma, W.-Y.: Web object retrieval, in *Proc. WWW 2007*, pp. 81–90 (2007)