情 報 学 基 礎　66－16
デジタル・ドキュメント 32－16
（２００２．３．１５）

文字列出現頻度比較による情報源間の類似性判定

佐藤 進也　原田 昌紀　風間 一洋
NTT 未来ねっと研究所
東京都武蔵野市緑町 3-9-11

Web サーバなどの情報源が持つリソース群中の文字列出現頻度を比較することで情報源間の類似性を判定する手法を提案する．これは，文書中の単語の使用頻度などを筆致を表す特徴量とし，その一致度から著者の同一性を判定する著者推定の手法を応用したものである．

本論文では，本手法を著者推定の一方法から導く過程を示す．さらに，本手法から導かれる情報源間の関係と，Web ディレクトリにおけるカテゴリの階層構造から導かれる情報源間の関係との整合性を調べ，本手法の妥当性を検証する．また，応用例として情報源の特徴語抽出について述べる．

Measuring Similarity among Information Sources by Comparing String Frequency Distributions

Shin-ya SATO, Masanori HARADA and Kazuhiro KAZAMA
NTT Network Innovation Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo

We propose a novel method for measuring similarity among information sources, such as web servers, by comparing distributions of string occurrence frequency in their resources. This approach is an analogue of the *literary detective*, which is to identify an author by comparing statistical characteristics of documents (e.g., word frequency distributions) that reflect authrs' writing styles. In this paper, we show how we have developed and validated the method. Similarity measured with this method is compared with that of derived from a Web directory service where information sources are classified and hierarchically arranged. We also describe a way to apply the similarity measuring method to selecting feature terms of information sources.

1　はじめに

文章の書き方には人それぞれのスタイルや個性がある．計量文献学 (あるいは計量言語学) では著者の個性を文章から計量的に抽出し著者推定に利用する研究が行われている [1]．これは，文の平均長や単語の使用頻度などを著者の個性を表す数量とし，その特徴量の一致度から著者の同一性を判定するものである．

我々はこの著者推定手法を情報源の特徴把握に応用した．ここで言う**情報源**とは，情報提供者が特定の目

的を持って運営している Web サーバの全体あるいは一部のことである．企業が運営している Web サーバや，Web ホスティング・サービスを利用して個人が作成・管理している Web ページ (の集まり) がその例である．

本論文では，著者推定応用の具体的方法とその検証結果および利用例を示す．以下，2節で著者推定の一手法を紹介する．さらに，そのアナロジーとして構成した，情報源の特徴量を算出する方法と特徴量を比較することで情報源間の類似性を判定する方法を述べる．この手法の妥当性を3節で検証する．4節では，本手法の応用例として情報源の特徴語抽出について述べる．さらに5節で関連研究を紹介する．

2　情報源の特徴抽出と類似性判定

2.1　Tankard の手法

著者推定のため文書の特徴量として，Tankard は文字の出現頻度に着目した [2]．文字 (この場合アルファベット) の集合を L，文書 D における文字 l の出現頻度を $f_D(l)$，D に含まれる文字の数を $|D|$ としたとき，Tankard が定義した D の特徴量は次の式で与えられる：

$$\left\{ \frac{1000}{|D|} f_D(l) \right\}_{l \in L} \tag{1}$$

式 (1) 中の $1000/|D|$ は，出現頻度 $f_D(l)$ を文書の大きさに関して正規化するためのもので，1,000 文字からなる文書を基準としている．

また，文書 D_1, D_2 の「隔たり」$\delta(D_1, D_2)$ を特徴量の L_1 ノルムで測っている：

$$\delta(D_1, D_2) = \sum_{l \in L} \left| \frac{1000}{|D_1|} f_{D_1}(l) - \frac{1000}{|D_2|} f_{D_2}(l) \right| \tag{2}$$

D_1，D_2 の著者がそれぞれ既知であり文書 D が著者がそのいずれかであるという仮定のもと，Tankard は $\delta(D_1, D)$ と $\delta(D_2, D)$ を比較することで著者を推定している．

2.2　提案手法

我々は，この Tankard の方法を以下のようにして情報源の特徴把握に応用した．

まず，文字列の集合 T に対して，情報源 w に属する Web ページで文字列 $t \in T$ を含むものの数を $F_w(t)$ とし，$F_w(T) = \sum_{t \in T} F_w(t)$ とする．式 (1) に対応する情報源の特徴量を次の式で定義する：

$$\left\{ \frac{F_w(t)}{F_w(T)} \right\}_{t \in T} \tag{3}$$

さらに，式 (2) にならって，2 つの情報源 w_1 と w_2 の「隔たり」を次式で定義する：

$$d(w_1, w_2) = \sum_{t \in T} \left| \frac{F_{w_1}(t)}{F_{w_1(T)}} - \frac{F_{w_2}(t)}{F_{w_2(T)}} \right| \tag{4}$$

著者推定のアナロジーにより導いた式 (4) は情報源間の意味的な近さ (類似性) を示していると考えられる．この仮説を次節で実験により確かめる．

3　検証実験

実験内容の説明に先立って，実験に用いた情報源の構成と文字列出現頻度の計算，文字列集合の構成方法をはじめの 3 節で説明する．次に，2 つの検証実験の内容と結果を3.4節と3.5節で示す．

3.1　情報源の構成

本実験では，Web ディレクトリ・サービス Open Directory[3] の `Top:World:Japanese` 配下に登録されているサイトで，URL が “/” で終わってるものを情報源の候補とした (図1左)[1]．そして，ある Web ページの URL(の表記文字列) がある情報源の URL に前方一致するとき，当該 Web ページはその情報源の一部であると仮定した．たとえば，情報源 w の URL が `http://somewhere.jp/` であるとき，URL が `http://somewhere.jp/foo.html` である Web ページは，この仮定により w の一部とみなされる．

統計的アプローチを用いている本手法の精度を向上させるため，実験の対象をある程度の大きさ (ページ数) をもった情報源に限った．大きさの判定には WWW 検索エンジン ODIN[4] の索引[2]を用いた．Open Directory から選んだ情報源の候補のうち，そこに属する Web ページが ODIN の索引中に 100 以上登録されている 2,101 サイトを実験対象として採用した．ここで得た情報源の集合を Ω とする．

ここで，「情報源」と「サイト」の関係 (言葉の用法) を整理しておく．「情報源」は「サイト」を，情

[1] 2001 年 8 月 5 日時点のデータ．
[2] 2001 年 8 月 30 日から 10 月 11 日にかけて収集した約 4229 万ページから作成したもの．

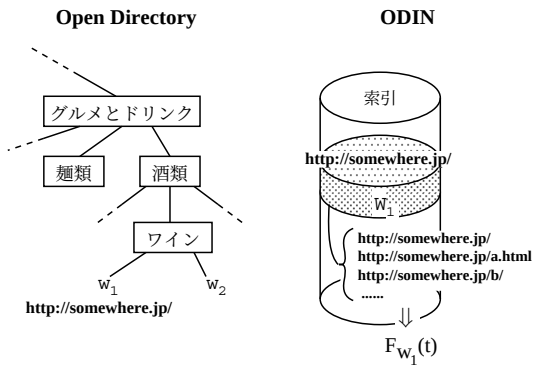

図 1: 実験に用いた 2 種類のデータ

報を利用する観点から捉えたときの呼び名である．それぞれが指し示す実体は同一のもの (ディレクトリの要素) であり，本論文ではこれらを適宜使い分ける．

3.2 文字列出現頻度の計算

情報源 w に属する Web ページで文字列 t を含むものの数 $F_w(t)$ を計算する手段として，ODIN の索引と全文検索モジュール Jerky[5] を利用した．$F_w(t)$ の値は，Jerky を用いて範囲を情報源 w に制限した文字列 t の検索を行い，その結果から算出する (図1右)．正確に言えば，ODIN の索引はすべての Web ページを網羅しているわけではなく，Jerky の索引作成方式 [5] に起因する検索特性や動作環境 (たとえばメモリ量) の制限により，すべての文字列の正確な出現頻度が調べられるわけではない．しかし，式 (3)，(4) は文字列生起の全体的な様子 (分布) によって情報源の特徴を把握 / 比較しようとするもので，個々の出現頻度の正確な値は重要ではない．本実験では改めて $F_w(t)$ を Jerky による検索結果の数と定義しなおし，以下議論をすすめる．

3.3 文字列集合 T の構成

文字列の集合 T としては次の 2 つを実験に用いた．1 つは Unicode[6] の 1 文字のみからなる文字列の集合 T_U である．T_U の要素で今回の実験で有用なもの，すなわち，すくなくとも 1 つの情報源 $w \in \Omega$ があって，$F_w(t)$ が正の値をとる文字列 t の数は 10,101 であった．

もう 1 つの集合は，EDR 日本語単語辞書 [7] に登録されている語で構成した T_E である．この辞書に登録されている約 19 万の名詞から今回の実験で有用な 10,101 語を T_E の要素として無作為に選んだ．

3.4 情報源の近傍

第一の実験として，d の値に基づいて情報源の近傍を構成し，その適切さを見た．情報源 w_0 の近傍は，$\Omega - \{w_0\}$ の要素 w を $d(w_0, w)$ に関して昇順にならべたときの最初の 10 サイトとした．

w_0 としてコンパックコンピュータ株式会社のサイト `http://www.compaq.co.jp/`，文字列集合として T_U を用いた結果を表1に示す．表において，各情報源を，その URL の `http://` 以降の文字列で示した．得られた情報源はすべて電子機器に関係する企業

表 1: `www.compaq.co.jp/` の近傍 (T_U)

情報源 w	$d(w_0, w)$
`www.panasonic.co.jp/`	0.40760
`www.sharp.co.jp/`	0.41308
`www.yamaha.co.jp/`	0.42717
`www.roland.co.jp/`	0.44296
`www.pioneer.co.jp/`	0.45833
`www.sanyo.co.jp/`	0.45951
`www.dospara.co.jp/`	0.46097
`www.hitachi.co.jp/`	0.46159
`www.ips.co.jp/`	0.46884
`www.fujixerox.co.jp/`	0.47256

のサイトであり，PC ベンダーであるコンパック社のサイトと関連のある情報源が収集できている．

一方，T_U の代わりに文字列集合として T_E を用いた結果が表2である．

表 2: `www.compaq.co.jp/` の近傍 (T_E)

情報源 w	$d(w_0, w)$
`www.apple.co.jp/`	0.68104
`www.sun.co.jp/`	0.70271
`www.hitachi.co.jp/`	0.70903
`www.unisys.co.jp/`	0.79022
`www.microsoft.com/japan/`	0.79314
`www.sw.nec.co.jp/`	0.79371
`www.sybase.co.jp/`	0.81404
`www.nec.co.jp/`	0.84306
`www.fujixerox.co.jp/`	0.84513
`www.panasonic.co.jp/`	0.86028

この場合，計算機関連企業のサイトが主な構成要素であり，T_U の場合よりも関連性の高い情

報源が収集できている．ちなみに，1位であった www.apple.co.jp/ は T_U を用いた場合には13位であった．

3.5 ディレクトリ構造との比較

次に，d による類似性判定の妥当性をさらに客観的に検証するため，ディレクトリ構造を利用して類似性の尺度を構成し，d との比較を行った．

3.5.1 ディレクトリ構造に基づく類似性尺度の構成

Web ディレクトリでは，カテゴリ間に階層的な構造が与えられている．また，異なる階層にあるカテゴリを同一視する「エイリアス」と呼ばれる関係が定義されている (図2)．たとえば，Open Directory では「アート: 映画: 俳優」は「レクリエーション: 有名人: 俳優」のエイリアスとして定義されている．

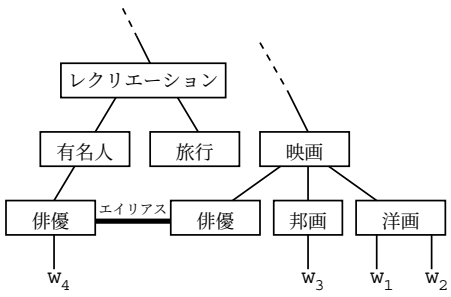

図 2: ディレクトリ構造と情報源間の関係

カテゴリをノードとし階層やエイリアスの関係をエッジとするグラフ構造と考えたとき，関連性の高いカテゴリはこのグラフ内で互いに近接した位置にあると考えられる．この関係を利用して，2つの情報源の類似性をそれらが属しているカテゴリの近さで測ることにする．その具体的な計算手順は以下の通りである．

まず，エッジのうち階層関係に対応するものの重みを1とし，エイリアスに対応するものの重みを0とする．この重み付きグラフにおいて，2つの情報源 w_1，w_2 それぞれが属するカテゴリ間の最短経路長を計算する [8]．この値を $h(w_1, w_2)$ と書く．図2において，$h(w_1, w_2)$, $h(w_1, w_3)$ はそれぞれ 0, 2 であり，$h(w_1, w_4)$ はエイリアスをたどることで 2 となる．情報源が複数のカテゴリに属している場合には，すべてのカテゴリの組み合わせについて最短経路長を計算し，その中で最小のものを h とする．

3.5.2 類似性尺度との比較実験

この h を類似性判定の基準とし，d をその値と比較する実験を行った．

まず，計算量を減らすため Ω から情報源をサンプリングした．Open Directory のカテゴリから10以上の情報源を含むものを無作為に20選び，それらに属している情報源を実験に用いた．ここで得られた情報源の集合を Ψ とする．Ψ の要素数はは287であった．Ψ から異なる任意の2つを取出して得られる 41,041 の組み合わせ (w_i, w_j) に対して $h(w_i, w_j)$ と $d(w_i, w_j)$ の関連性を調べた．その結果，個々の値の相関係数は 0.14165 と正ながら低い値であった．一方，h の値と d の分布との間にははっきりとした相関が見られた．

$h(w_i, w_j) = k$ という条件下での d の値の累積確率分布 $P(k, x)$，すなわち x に対して $d(w_i, w_j) < x$ が満たされる場合の数の全体における割合は次の式で与えられる[3]：

$$P(k,x) = \frac{|\{(w_i,w_j)|\ d(w_i,w_j) < x,\ (w_i,w_j) \in H_k\}|}{|H_k|},$$

$$H_k = \{(w_i,w_j)|\ h(w_i,w_j) = k,\ w_i,w_j \in \Psi,\ w_i \neq w_j\}$$

$P(k, x)$ の変化をいくつかの k ごとに示したものが図3である．$k = 1$ の場合は，得られたデータ数が少く分布の様子を把握し難いので図から除いた．(A) は d を計算する際に文字列集合として T_U を用いた場合であり，(B) は T_E を用いた場合である．

これらのグラフでは，分布を示す曲線が左側に寄っているほど d が小さな値を高い確率でとることを意味している．この性質により，h と d の関係が，k の値と対応する曲線の位置との関係から導かれる．図では，T_U の場合に $k = 2$ の曲線が $k = 0$ の左側に位置している以外，k の増加に伴い曲線の位置が右に移っている．このことから，d は情報源の類似性に関して尺度 h と似通った評価を与えることがわかる．

2つのグラフを比べると，(B) の方が k の値と曲線位置の相関が高いうえ曲線の間隔が大きい．これは，

[3] 集合 S に対して $|S|$ はその要素数を表す．

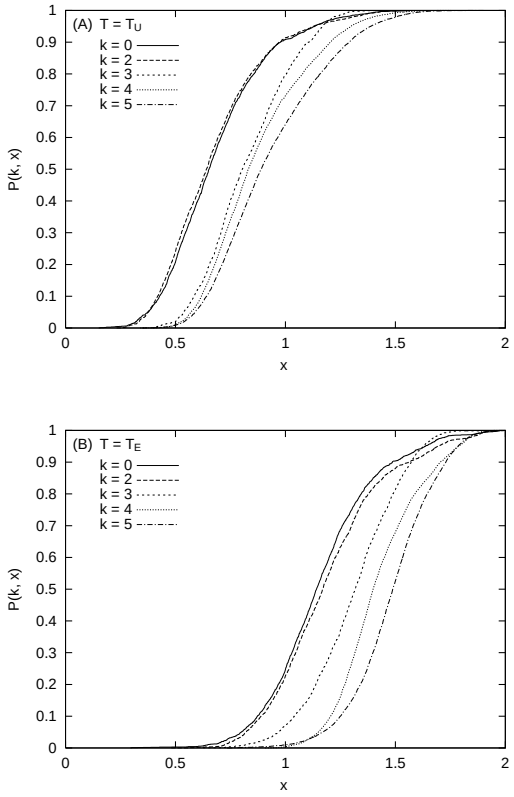

図 3: d の累積確率分布

T_E を用いた方が類似性判定の精度が高いことを示しており，3.4節の結果とも一致する．

3.5.3 低相関の原因

前節の解析の結果は，h と d の間には全体的に相関がないのではなく，相関を妨げている情報源が一部に存在していることを示している．その原因を 2 つの場合に分けて調べた．

h は小さな値，d は大きい値をとる場合： 2 つの情報源が同様な内容を扱っていても，一方が特異な文字列出現分布を持つ場合には d の値が大きくなる．特異な分布はデータの片寄りの現れであり，情報源がある特定のデータを多く持っている場合にこのような状況が発生する．多くの住所データを含む観光地の宿泊施設情報を提供しているサイトがその一例である．これは，本手法の効果が情報の表現に依存していることを示している．

h は大きな値，d は小さな値をとる場合：この原因の一つに，ディレクトリの不完全さがあると考えられる．情報源間の関係がすべてディレクトリに記述されているとは限らず，その結果として h が大きな値をとることがある．例えば，地方自治体のサイトとその地方にある教育機関のサイトには「同一の地方に所在する」という関係があるが，それがディレクトリに記述されていない (具体的には，前者は地域情報のカテゴリに属しているが，後者は教育機関のカテゴリのみに属している) ものがある．

もう一つの原因は本実験における情報源の構成方法にある．そもそも，情報源としては何らかの主題を持って構成されている Web ページ群を想定しており，主題をもとに構成されているゆえ，そこには何らかの (統計的) 特徴が認められる，という仮説に本手法は基づいている．しかし，3.4で述べた情報源の構成方法では，内容的に関連性の低い Web ページが情報源の一部となってしまうことがある．インターネット・サービス・プロバイダ (ISP)[4]のカテゴリに属している情報源が，ウェブ・ホスティング・サービスを利用している多数の個人サイトを含んでいる場合がその例である．この場合，情報源の特徴が適切に抽出されない可能性がある．実際， ISP のサイトと，ゲームメーカーや新聞社のサイトとの間で d が低い値を示している．

4 応用例

類似度 d は，情報源間の関係把握や情報源自身の特徴抽出のために利用できる．前者の例としては，近傍の構成 (3.4節) やクラスタ解析による分類 [9] が挙げられる．本節では後者の例として，情報源の特徴語を抽出する方法について述べる．

4.1 語の特定性と情報源の類似性の関係

文書からその内容を表す語 (特徴語) を抽出する方法として tf·idf 法 [10] がよく使われている．これは，当該文書内で数多く言及され，さらに特定性のあるもの，すなわち文書集合の中で少数の文書にのみに出現する語を選び出すというアプローチである．文書集合に類似性の尺度を導入したとき， tf·idf 法のアプローチは「ある文書 D との類似性が高い文書には高頻度

[4] Open Directory の「アクセス・プロバイダ」のカテゴリに対応するもの．

で出現し，類似性の低い文書には低頻度で出現する語は D の特徴をよく表している」と言い換えることができる．この考え方を情報源と類似度 d に適用し，情報源の特徴語抽出を試みた．

まず，基準となる情報源 w_0 を決め，別の情報源 w を w_0 に近い ($d(w_0, w)$ が小さい) ものから遠い ($d(w_0, w)$ が大きい) ものに変えていったとき，語 t の正規化された出現頻度 $F_w(t)/F_w(T)$ がどのように変化するかを見た．w_0 として `www.compaq.co.jp/` を，t として特定性の異なる “システム” と “ホームページ” を選んだ場合の分布を図4に示す．全体の傾向を捉え易くするため，それぞれのグラフで横軸方向に隣り合う 2 点を直線で結んだ．また縦軸方向のスケールは 2 グラフ間であえて揃えていない．

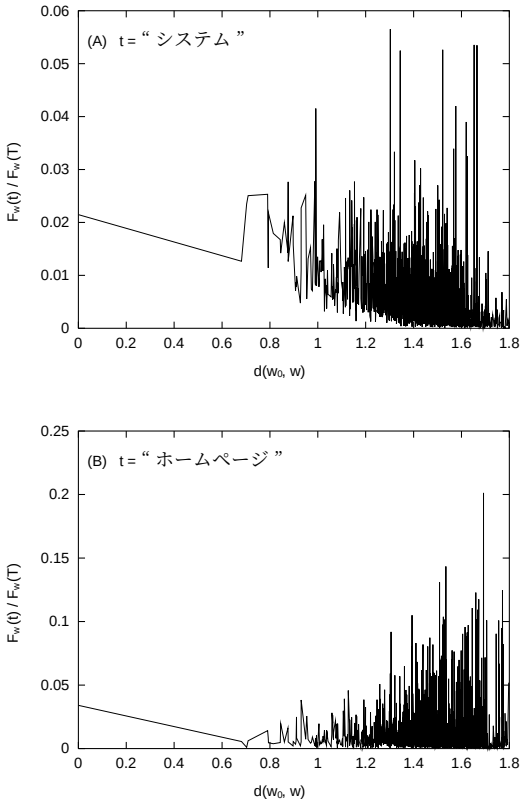

図 4: $d(w_0, w)$ と $F_w(t)/F_w(T)$ の相関

情報源 w_0 と関連があり，特定性が比較的高いと思われる “システム” のグラフ (A) では，全体的に右下がりの傾向 (負の相関) が認められる．一方，“システム” に比べて特定性の低い “ホームページ” のグラフ (B) ではそのような相関は認められない．この結果は，$d(w_0, w)$ と $F_w(t)/F_w(T)$ の相関が w_0 の特徴語 t を選ぶ基準となり得ることを示唆している．

4.2 相関係数に基づく特徴語抽出

そこで，相関の様子を把握するための数量として $d(w_0, w)$ と $F_w(t)/F_w(T)$ の相関係数 $r(t)$ を採用し，その値に基づく特徴語抽出を試みた．具体的には，与えられた語の集合に対して，その要素 t を $r(t)$ の値により昇順に並べ，上位 40 語を (その集合の中で) 特徴的な語として選択した．これを，単純に出現頻度 $F_{w_0}(t)$ の大きい順に選んだ場合と比較し本手法の有効性を確認した．なお，$r(t)$ の値は次の式で与えられる：

$$r(t) = \frac{\sum_{w\in\Omega} \left(d(w_0,w) - \overline{d}\right)\left(\frac{F_w(t)}{F_w(T)} - \overline{F}\right)}{\sqrt{\sum_{w\in\Omega}\left(d(w_0,w)-\overline{d}\right)^2}\sqrt{\sum_{w\in\Omega}\left(\frac{F_w(t)}{F_w(T)}-\overline{F}\right)^2}},$$

$$\overline{d} = \frac{1}{|\Omega|}\sum_{w\in\Omega} d(w_0,w), \quad \overline{F} = \frac{1}{|\Omega|}\sum_{w\in\Omega}\frac{F_w(t)}{F_w(T)}$$

語の集合として，ある 1 日間に ODIN に寄せられた検索語[5]を用いた場合の結果が表3である．出現頻度を用いた場合に上位に現れる “ホームページ”，“利用”，“情報” といった特定性の低い語が，相関係数を用いた場合には除かれている．その一方で，“compaq”，“コンピュータ”，“ソフトウェア” といった高頻度でかつ特徴的な語は相関係数を用いた場合にも抽出されている．この結果は，特徴語抽出の尺度として相関係数が有用であることを示している．

5 関連研究

ここでは本研究に関連する二つの話題として，分布比較の手段，分散検索における情報群の特徴抽出，Web ページ間の関連性把握に関する研究をとりあげる．

5.1 分布比較の手段

文字列出現頻度分布間の隔たりを測る手段には，本手法で用いた L_1 ノルムの他に，文書間の類似性を測るためによく用いられているベクトル空間モデル [10]

[5] 2001 年 11 月 1 日に寄せられた 9,008 語.

表 3: 2 つの方法で抽出された上位 40 語

出現頻度	ホームページ, 利用, compaq, システム, サービス, ソフト, 情報, Windows, リ, ソフトウェア, ¥して, コンピュータ, NT, ダウンロード, モデル, 方法, 価格, ネットワーク, データ, 仕様, Microsoft, メモリ, PC, ディスク, 日本, 2, 概要, プログラ, インストール, 一覧, 株式会社, ネット, 表示, と, 処理, Server, アクセス, 画面, for, /.jp
相関係数	インストール方法 windows2000, タグ付きコマンド, ip 検索, 障害管理, ERP データベース, アダプタ ネットワーク, 障害予防監視, metaframe, NT, ソフト, ソフトウェア, raid, Windows, アプライアンスサーバ, TCO, システム, compaq, インストール, 筆グルメ, プロトコル, LAN, Microsoft, USB, SCSI, CPU, システム構築, メモリ, コンピュータ, トランザクション, 構成済, フォーマット, コネクタ, サーバを動作, ツール, ATI, Server, サービス, パソコン, PC, プリンタ

の適用が考えられる．具体的には，文字列出現頻度分布をベクトルとみなし，2 つのベクトルのなす角度を分布間の隔たりとする．

この 2 つの手段を用いて，いくつかの情報源の近傍をそれぞれ構成したところ，ベクトル空間モデルの場合には関連性の低い情報源の混入が若干認められた．この範囲では L_1 ノルムがやや優れているという結果が得られたが，総合的な評価のためにはさらに詳しい比較が必要である．また，この他の選択肢 (松浦らによる *sim*[11] など) との比較も今後の課題である．

5.2 分散検索におけるメタデータ

情報の集まり (コレクション) からの特徴抽出は，分散検索の研究における重要な課題である．多くの分散検索システムでは，分散管理されているコレクションごとにその特徴を表すデータ (メタデータ) を作成し，検索要求に適合するコレクションの効率的な絞り込みに用いている．

Harvest[12] のメタデータは，文書の著者やサイズといった属性とその値の組 (attribute-value) からなる．その記述や情報交換には Summary Object Interchange Format(SOIF) という形式を用いている．Ultraseek[13] では情報の内容を要約した識別特徴 (fingerprint) を作成している．また，CAFE[14] では文書の n-gram 統計をもとにメタデータを作成している．これらは，どのような特徴をもった情報がどのコレクションに存在するかという一種の索引をメタデータとして構成するもので，情報源全体の特徴を捉えようとする本手法と基本的なアプローチが異なる．

情報源の全体像を捉えているものとしては OBIWAN[15] がある．OBIWAN では知的エージェントが情報源の特徴を抽出し，それをオントロジーとして記述する．OBIWAN のアプローチと比べると，本手法で得られる特徴は詳しさや具体性に欠るが，簡易な処理で抽出できる点が優れている．

5.3 情報源の構成と Web ページ間の関係把握

高精度な特徴抽出のためには，まず適切な情報源の構成が必要である．しかし，本論文で示した方法ではこの要件が必ずしも満たされておらず，改善の余地を残している．

今回の我々の実験では，Web ディレクトリから取り出した URL と前方一致する Web ページを集めて情報源を構成した (3.1節)．この方法では，情報源全体の意味的なまとまりは必ずしも保証されない．実際，多種多様なページが混ざり合い情報源として適切でないものが構成されてしまう場合もあった (3.5.3節)．情報源の意味的な統一性を保証するには，Web ページ間の関係を調べ関連性の弱いものは除外する，などといった処理が必要である．

Web ページ間の関係を調べる手段としては，ハイパーテキストのリンク構造解析 (リンク解析)，内容の比較，ユーザの情報利用パタン解析がある [16]．特に，リンク解析は最近盛んに研究されているテーマである．これは，リンクによる結び付き方から Web ページ間の関係を推測するもので，検索結果のランク付け [17, 18]，コミュニティの発見 [19] や Web ディレクトリの拡張 [20] などの成果がある．さらに，リンク解析は，意味的にまとまりのあるページ集合を構成する手立てとして用いられており [21, 22]，情報源の構成方法を改善する手段としても期待できる．

6 むすび

情報源を文字列の出現頻度分布により特徴付けし，その L_1 ノルムで情報源間の類似度を測る方法を提案した．また，本手法の妥当性を，Web ディレクトリ・サービスの構造から導かれる類似性の尺度と比較することで確認した．低い処理コストに対して十分な効果が得られるところに本手法の特徴がある．

本手法には，4節で述べた特徴語抽出以外にも様々

な応用が考えられる．特に，それぞれの情報源に検索機能を持たせ，本手法により抽出した特徴をメタデータ (5.2節) として用いて分散検索を実現した場合，既存のシステムとは違ったスタイルの情報ナビゲーションを提供できる可能性がある．

今後は，実験の結果明らかになった問題点を改善しつつ，分散検索への適用を検討していく予定である．

謝辞

Web ディレクトリの運用とデータ提供に対して，Open Directory Project に感謝の意を表します．

参考文献

[1] 村上： “著者はだれか？ 計量文献学への招待 1”, 数学セミナー, 11 月号, pp.55–59, 1988. (1989 年 3 月号までの 5 回連載)

[2] J. Tankard： “The Literrary Detective,” BYTE, Vol. 11, No. 2, pp. 231–238, 1986.

[3] Open Directory Project： “Open Directory,” http:// dmoz.org/.

[4] ODIN：http://odin.ingrid.org/.

[5] 原田, 他： “Unicode を用いた N-gram 索引の一実現方式とその評価”, 情報処理学会研究会報告 2000-FI-57-17, 2000.

[6] The Unicode Consortium： “The Unicode Standard, Version 2.0,” 1996.

[7] 株式会社 日本電子化辞書研究所：EDR 電子化辞書 1.5 版, 1996.

[8] Dijkstra E.W.： “A note on two problems in connection with graphs,” Numerical Mathematics, (1), pp. 269–271, 1959.

[9] M. R. Anderberg： “Cluster Analysis for Applications,” Academic Press, 1973.

[10] G. Salton, M. J. McGill： “Introduction to Modern Information Retrieval,” McGraw-Hill, 1983.

[11] 松浦, 金田： “n-gram 分布を用いた近代日本語小説文の著者推定”, 計量国語学, Vol.22, No.6, pp.225–238, 2000.

[12] C. M. Bowman, et. al., “The Harvest Information Discovery and Access System,” Computer Networks and ISDN Systems, Vol.28, pp.119–125, 1995.

[13] S. Kirsch： “Infoseek’s approach to distributed search,” Distributed Indexing/Searching Workshop, 1996. http://www.w3.org/Search/9605-Indexing-Workshop/Papers/Kirsch@Infoseek.html.

[14] G. Crowder and C. Nicholas： “Resource selection in CAFE: An architecture for network information retrieval ,” In Proceedings of the Network Information Retrieval Workshop, SIGIR 96, 1996.

[15] X. Zhu, et. al,： “Ontology-Based Web Site Mapping for Information Exploration,” in Proceedings of the 8th International Conference on Information and Knowledge Management, pp. 188–194, 1999.

[16] C. Chen： “Structuring and Visualising the WWW by Generalised Similarity Analysis,” in Proceedings of the 8th International ACM Conference on Hypertext and Hypermedia, 1997.

[17] J. Kleinberg, “Authoritative sources in a hyperlinked environment,” Proc. of the 9th ACM-SIAM Symposium on Discrete Algorithms, pp. 668–677, 1998.

[18] S. Brin and L. Page, “The Anatomy of a Large-Scale Hypertextual Web Search Engine,” Proc. of the 7th International World Wide Web Conference, 1998.

[19] 豊田： “WWW における関連コミュニティ群の発見”, 情報処理学会研究会報告, 2000-DBS-122-40, 2000.

[20] 原田, 他： “参照共起分析の Web ディレクトリへの適用”, 情報処理学会研究会報告, 2001-FI-61-7, 2001.

[21] L. Terveen, et. al.,： “Constructing, Organizing, and Visualizing Collection of Topically Related Web Resources,” ACM Trans. Coumputer-Human Interaction, Vol. 6, No. 1, pp. 67–94, 1999.

[22] 原田, 他： “WWW ページ間の階層構造の推定と検索システムへの応用”, 情報処理学会研究会報告 99-DBS-118-14, 1999.