社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

人は Web 上の同姓同名人物をどのように判別しているのか

三宅 悠生† 村上 晴美‡

†‡大阪市立大学大学院創造都市研究科 〒558-8585 大阪市住吉区杉本 3-3-138
E-mail: †dragonball19850913@gmail.com ‡harumi@media.osaka-cu.ac.jp

あらまし 本研究は，Web 上の人物検索に寄与する知見を得るために，人間による同姓同名人物の判別過程を明らかにすることを目的とする．20 の人名をクエリとして検索エンジンで得た上位 20 件のデータを被験者に分離させ，質問紙，観察，インタビューにより分析した．既知人名と未知人名でサイト閲覧をする割合が異なること，人物を識別するためにはキーワードと職業と作品（実在人物の場合は該当人物の著作物，架空人物の場合は該当人物の登場する作品）が重要であることなどがわかった．実験結果に基づき，同姓同名人物の分離モデルと知識構造モデルを提案した．

キーワード Web 人物検索，同姓同名人物分離，同姓同名人物の分離モデル，知識構造モデル

How Do Humans Identify Different People with Identical Names on the Web?

Yuki MIYAKE† and Harumi MURAKAMI‡

†‡Graduate School for Creative Cities, Osaka City University 3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585 Japan
E-mail: †dragonball19850913@gmail.com ‡harumi@media.osaka-cu.ac.jp

Abstract This research investigates how humans identify different people with identical names on the Web to obtain knowledge that is helpful for searching for people on the Web. We asked subjects to classify 20 Web people-search results for each of 20 Japanese person names (i.e., 20 pages obtained for each person name) and analyzed their decision processes by questionnaire, observation, and interview. We found that the rate of browsing sites differs between known and unknown names and that keywords, vocations, and works (when the person is a real person, works are those he or she has made, when the person is a fictional character, works are those in which he or she appears) are important for distinguishing individuals. We propose a model for distinguishing individuals and a knowledge-structure model based on the experiment's results.

Keyword Web People Search, Person Name Disambiguation, Distinguishing Individual Model, Knowledge-Structure Model

1. はじめに

ネット上で情報発信する人々が増加するに伴い，Web 上の人物検索においては，同姓同名人物の識別が重要な課題となっている[1]．これまでに，同姓同名人物の自動分離の研究が数多く行われてきたが，技術的に難しい問題であり，完全な自動化は難しい．

本研究は，Web 上の人物検索に寄与する知見を得るために，人間による同姓同名人物の判別過程を明らかにすることを目的とする．

本稿では，2 節で実験の方法，3 節で結果と考察を述べる．4 節では実験に基づいて得られた，判別しやすい人物の特徴を述べ，5 節では人間による同姓同名人物の分離モデルと同姓同名人物識別に利用する知識構造モデルを提案する．6 節で関連研究と比較して議論を行う．

2. 方法

被験者は 14 名（男性 9 名，女性 5 名，平均年齢 25 歳）である．学生が 11 名，社会人が 3 名である．

先行研究で使用された 20 の人名[2]をクエリとして Yahoo!API で上位 20 件 の HTML，タイトル，スニペット，URL のデータを取得して実験用サイトを構築した．20 の人名に 2 人の被験者を割り当てた．すなわち，20 人名×2 人=40 件のデータとなる．

被験者にはまず質問紙により，検索に使われた人名の人物を知っているかどうかについて回答させた．

次に，実験用サイトの人名検索結果の一覧画面を見せて，20 件の分離を行わせた．一覧画面にはタイトル，スニペット，URL が表示されている．タイトルを選択すると，該当ページが表示できる．1 件目を 1 人目（1 番）とし，残りの 19 件について人物に番号をつけて記

載させた．実験中は録画を行った．

分離終了後，質問紙を用いて，以下の質問に答えさせた．

(a)　タイトル，スニペット，URL，Web 文書の中から，人物を分離する際に参考になったキーワードを重要度の高いと感じる順に 1〜10 個以内で列挙して下さい．

(b)　分離された人物に対して，その人物を最も特徴付けると思うキーワードを 1 つ記入して下さい．

実験後にインタビューを行った．

3. 結果と考察

3.1. 質問紙

3.1.1. 正解率

人手により正解データを作成した．全部で 58 人物存在した．被験者が分離した 1 ページ毎に正解データと比較したところ，被験者の正解率は 79%となった（表 1）．

各人名別では，被験者 2 人とも正解率が低い人名が幾つか見られた．野村紀子，木下和彦，田中克己の人名に共通しているのが 6 名以上の同姓同名人物が存在していることであった．

表 1：被験者の正解率

氏名	人物数	正解率 被験者1	正解率 被験者2	氏名	人物数	正解率 被験者1	正解率 被験者2
江川卓	2	0.90	1.00	野村紀子	6	0.55	0.25
三浦麻子	4	1.00	0.90	和田英一	2	0.90	0.90
山岡士郎	4	0.85	0.85	伊庭幸人	1	1.00	0.65
上田次郎	5	1.00	0.75	栗原はるみ	1	1.00	1.00
新垣紀子	2	0.70	0.65	五斗進	1	1.00	1.00
竹内郁雄	1	0.90	1.00	五嶋みどり	1	1.00	1.00
中村紘子	1	0.95	1.00	野間佐和子	1	0.65	0.95
田中克己	8	0.70	0.55	畑村洋太郎	1	0.95	1.00
菱沼聖子	6	0.85	0.65	福原愛	1	0.50	1.00
木下和彦	9	0.50	0.50	水野晴郎	1	0.15	0.90
合計					58	0.79	

同姓同名人物が 6 名以上いるかいないかに分けてマンホイットニーの U 検定で分析したところ，有意差がみられた（$p<.01$）．同姓同名人物が 6 名以上存在する人名では正解率が下がると言える．

3.1.2. 既知人名と未知人名

質問紙 40 件の結果の中，7 件（18%）が既知の人名という回答であった．7 件は江川卓 2 件，山岡士郎 1 件，中村紘子 1 件，五嶋みどり 1 件，福原愛 2 件である．実在人物については直接知っているわけではなくメディア等で知っており，漫画の登場人物である架空人物については漫画で知っていた．すべての被験者において直接知っている人物はいなかった．

既知人名でサイト閲覧する割合の平均が 34%であるのに対して，未知人名では 61%であった．既知人名の場合，被験者は少なくとも一人の人物に関するある程度の事前知識を持っているため，スニペットとタイトルで判別できる場合が多いと考える．

3.1.3. 識別キーワードと特徴キーワード

記入されたキーワードについて集計したところ，質問(a) が 329 個，質問(b) が 126 個であった．前者を識別キーワード，後者を特徴キーワードと呼ぶ．

これらのキーワードを著者らによりカテゴリに分類した．図 1 に識別キーワード，図 2 に特徴キーワードのカテゴリへの分類結果を示す．

図 1, 2 より，識別キーワード，特徴キーワードともに，上位 3 カテゴリが，キーワード，職業，作品であることがわかる．識別キーワードではキーワード 56%，職業 29% 作品 8%という割合で，特徴キーワードではキーワード 44%，職業 44%，作品 6%という割合であった．

なお，キーワードカテゴリは，被験者が記入した様々なジャンルの概念（多くは単名詞あるいは複合名詞）を表しており，他のカテゴリ（職業，作品，経歴，趣味，URL，画像）に該当しないものをまとめている．所属は職業カテゴリに含めている．

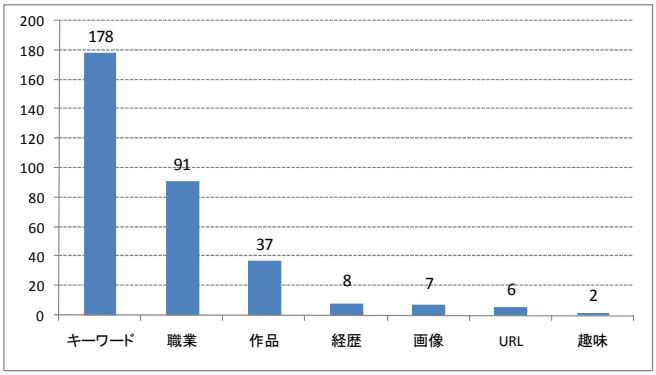

図 1：識別キーワードの分類結果

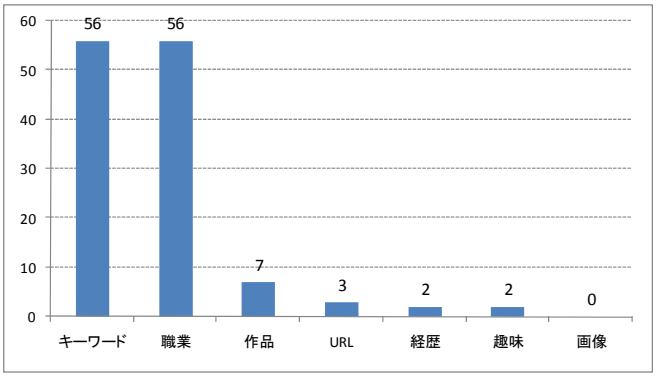

図 2：特徴キーワードの分類結果

作品は，主に人物の著作である著書や作品であるが，本実験には TV ドラマや漫画等に登場する架空人物が含まれており，架空人物の登場する作品である場合がある．そこで，実在人物と架空人物に分けて上位 3 カテゴリの集計を行った．結果を表 2 に示す．実在人物は 54 人，架空人物は 4 人存在した．

実在人物では，識別キーワードとして「キーワード」が突出しているが，特徴キーワードとして「職業」が「キーワード」よりわずかに多い．架空人物では識別も特徴も 1 位は「作品」となっている．

表 2：実在人物と架空人物

		キーワード	職業	作品
実在人物	識別	166	88	24
	特徴	54	56	2
架空人物	識別	12	3	13
	特徴	2	0	5

以上より，人物の判別によく使われるのは「キーワード」であり，次に「職業」「作品」と考えられる．架空人物に限ると，該当人物の登場する「作品」となる．人物を特徴付けるものは，実在人物の場合は「職業」と「キーワード」であり，架空人物の場合は「作品」と「キーワード」であることがわかる．

3.2. プロトコル分析

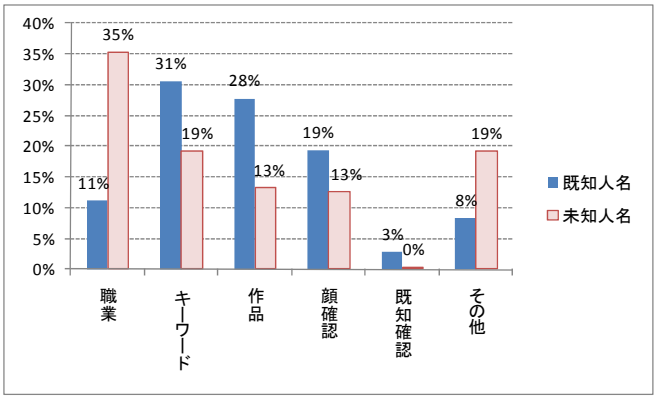

図 3：プロトコル分析結果

録画から，人物を判別する際の発話を抽出し，既知人名と未知人名に分けてカテゴリに分類した（図 3）．全体的に，職業，キーワード，作品，顔確認の順番に発話が多かった．既知人名では，キーワード，作品，顔確認，職業の順番で，未知人名では，職業，キーワード，作品，顔確認の順番であった．

質問紙では回答の少なかった顔画像の重要性がうかがえた．今回の実験の場合，既知人名は有名人であると考えられ，有名人においてはキーワードや作品，顔画像が重要であると言えるだろう．

3.3. インタビュー

「人物を判別する際に未知の人名では事前知識がないためにサイト閲覧を行った」という回答や「判別の際には何かしらのイメージをしている」という回答を得た．イメージしている内容としてまず顔等の画像があげられた．人物を被験者が判別するための条件を回答してもらったところ，「顔，職業，経歴，実績，所属に関する情報が発見できるかどうかである」という意見が得られた．中には「著名な業績が 1 つ見つかれば人物を特徴付けられる」という回答や「顔画像からその人物に関しての多くの情報を読み取れるから重要である」という回答も得られた．

4. 判別しやすい人物の特徴

未知人物において正解率の高いデータを分析し，判別しやすい人物の特徴として以下の項目を洗い出した．(1) 顔画像, (2) テキスト（内容）, (3) 情報量, (4) 分けやすさの 4 種類に大別できる．

(1) 顔画像

① 人物毎の 3 件目以内に顔画像がサイト内に記載

(2) テキスト（内容）

② 職業（所属を含む）または経歴がスニペットまたはサイトに記載

③ 業績がサイトに記載

④ 職業または経歴または業績に関連するキーワードがタイトル，スニペットまたはサイト内に記載

⑤ Wikipedia のページが存在

(3) 情報量

⑥ 同一人物のページが 3 件以上存在

(4) 分けやすさ

⑦ 同姓同名人物が 6 名未満

⑧ 同姓同名に類似する職業や分野の人物が存在しない

ただし，ここで経歴とは，生年などの時情報を含む出来事であり，業績とは，主として職業上の成果である．

テキスト（内容）に関しては，職業と経歴と業績が重要である．Wikipedia はこれらの情報が詳細に記述されていることが正解率を高めた主な要因であると考えるが，そのほかにも，情報量が多いことや有名人である場合が多いことも要因であると考える．

5. 同姓同名人物の判別モデル

実験で得られた知見を基に同姓同名人物の判別のモデルを考案した．図 4 では被験者が与えられた人名

検索結果一覧を同姓同名人物に分離する過程，図 5 では被験者が人物を判別する際に利用している知識をモデル化した．前者を同姓同名人物の分離モデル，後者を知識構造モデルと呼ぶ．

同姓同名人物の分離モデル（図 4）では，まず，タイトルやスニペットなどを見て，知識構造を参照して，人物が既知か未知かを判別し，既知の場合は人物を分離する．未知の場合や，既知であるが確認が必要な場合はサイトを閲覧する．サイト閲覧により，人物に関連する情報を取得し，人物に関連する知識構造の作成，追加，修正を行う．その後，人物を分離する．

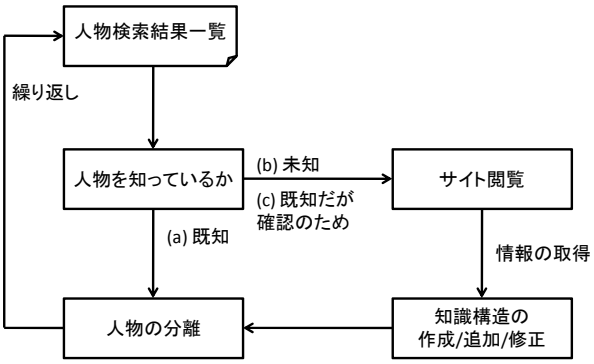

図 4：同姓人物の分離モデル

知識構造モデル（図 5）は，人物に関連する知識（知識構造 1）と Web サイトに関連する知識（知識構造 2）に大別できる．プロトコル分析やインタビュー調査の結果より，人物に関連する知識には参照の際の優先度があると考える．優先度 1 を顔画像，優先度 2 を内容とした．内容は，実在人物と架空人物に分け，実在人物の中では，職業，キーワード，著作物，経歴の順番とし，架空人物の中では，作品名，キーワード，職業の順番としている．作品は実在人物の場合著作物とし，架空人物の場合作品名とした．実在人物においては質問紙の集計はキーワードの方が多いが，プロトコル分析やインタビュー結果より職業を優先させている．また，実験結果より，キーワードの中から人名を抽出して別枠とした．

一方，Web サイトに関連する知識は URL，リンク先のページ，サイトの構造とした．

6. 関連研究

上田ら[3] は同姓同名人物の人間による識別を支援するために職業関連情報を抽出する手法を提案した．職業関連情報として職業，所属，著作に着目したが，本研究は実験によりその考えの裏付けを得るとともに，顔画像などの重要性を新たに確認している．

Yoshida ら[4] はブートストラッピングを用いた 2 段階クラスタリングを自動分離に適用し，1 段階目に固有表現，URL，複合キーワードを特徴としてクラスタリングし，2 段階目にブートストラッピングを用いてクラスタリングしている．本研究では，人間が人物の分離を行う場合に重要な特徴を示している．

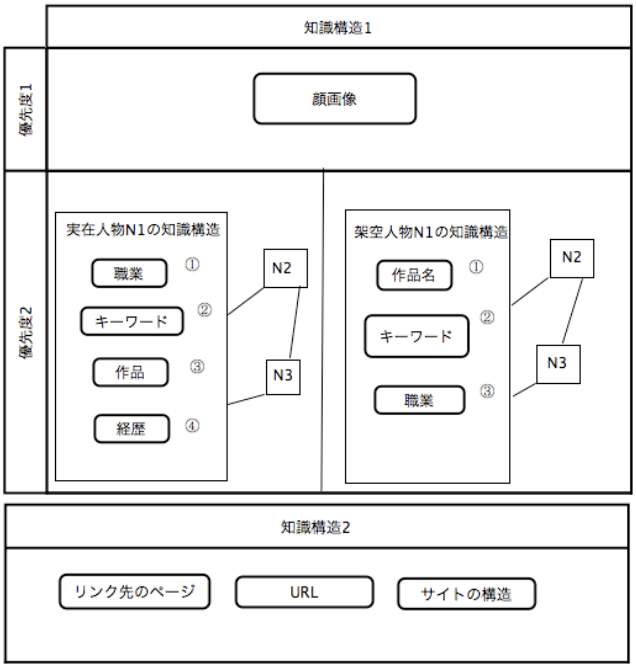

図 5：知識構造モデル

7. おわりに

Web 上の同姓同名人物を人がどのように判別しているのかを調べるための実験を行った．実験に基づき同姓同名人物の判別モデルを提案した．

文　献

[1] 関根聡，“Web 検索における人名の曖昧性解消技術の動向―同姓同名のクラスタリング―，” 情報処理，vol.49, no.5, pp.573-578，2008.

[2] 佐藤進也，風間一洋，福田健介，村上健一郎，“実世界指向 Web マイニングによる同姓同名人物の分離，”情報処理学会論文誌：データベース，vol. 46, no. SIG8, pp. 26‐ 36，2005.

[3] 上田洋，村上晴美，辰巳昭治，“Web 上の同姓同名人物識別のための職業関連情報の抽出，” システム制御情報学会論文誌，vol.22, no.6, pp.1-12, 2009.

[4] M. Yoshida, M. Ikeda, S. Ono, I. Sato, and H. Nakagawa, “Person Name Disambiguation by Bootstrapping, ” Proc. SIGIR’10., pp.10-17, Geneva, Switzerland, Jul. 2010.