WEBを用いた人物の別名抽出

本間 大輝† (honma@mi.ci.i.u-tokyo.ac.jp)
Danushka Bollegala†, 松尾 豊††, 石塚 満†

†東京大学大学院
††産業技術総合研究所

背景

- 検索エンジンの利用が増加

↓

- 検索性能向上が重要な課題

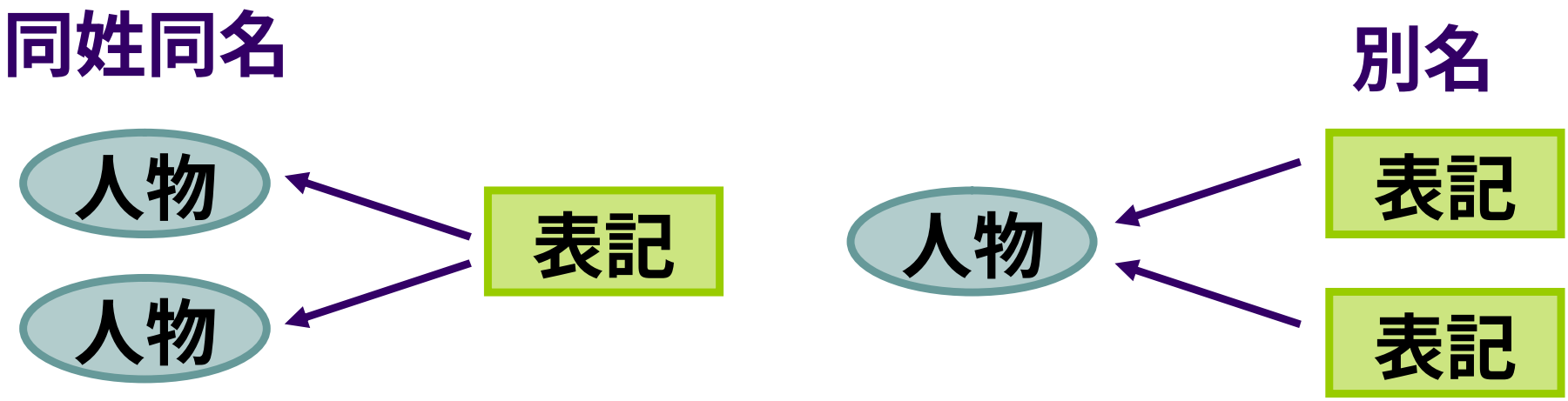
↓

全クエリの３０％を占める
人名の検索がとりわけ重要

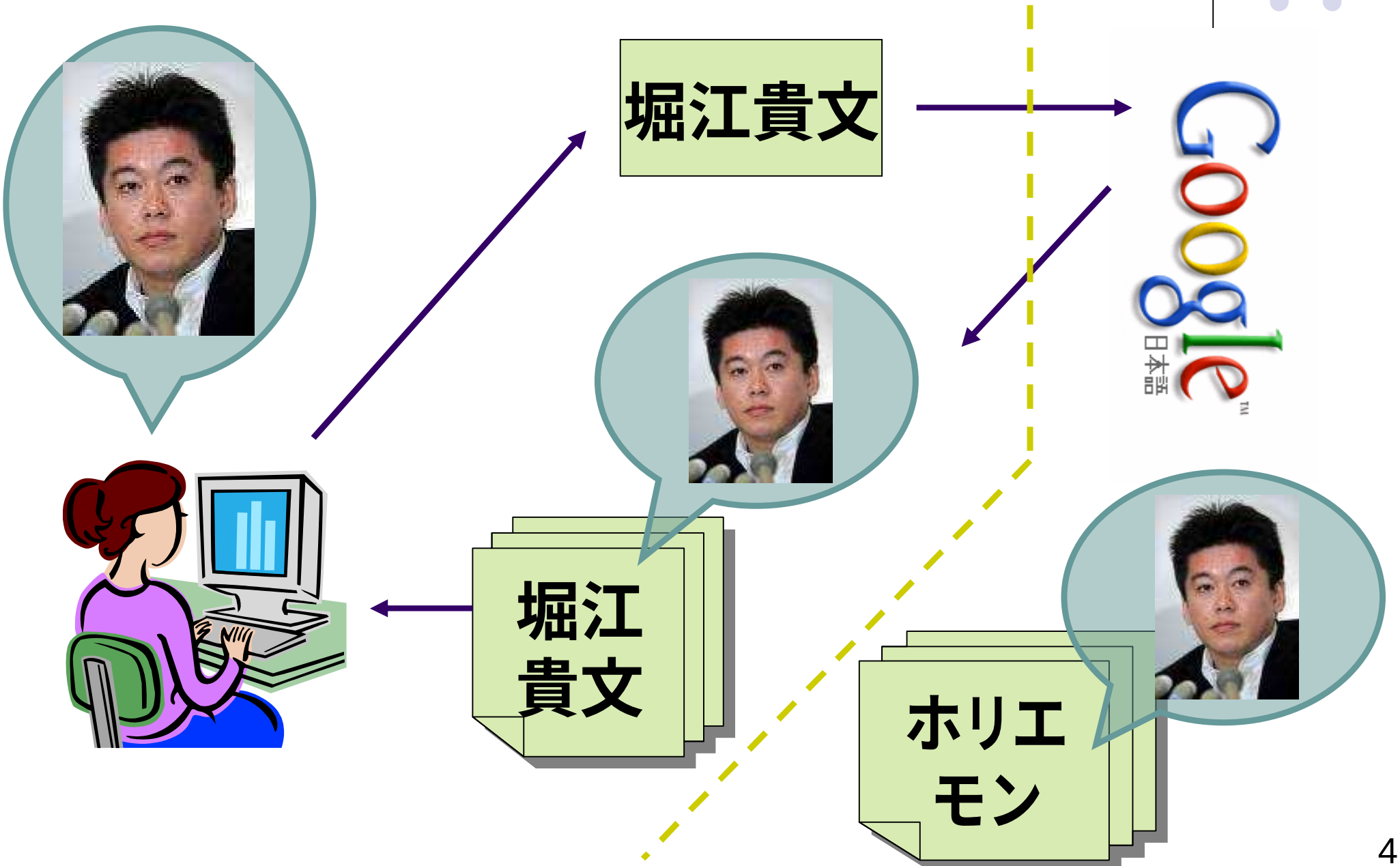
J. Artiles, et al. , ”A Testbed for People Searching Strategies in the WWW” , SIGIR05.

人名検索における問題

- 検索するためには，対象をキーワードで表現しなければならない．
- だが，実世界の人物と，その表記は必ずしも一対一には対応しない

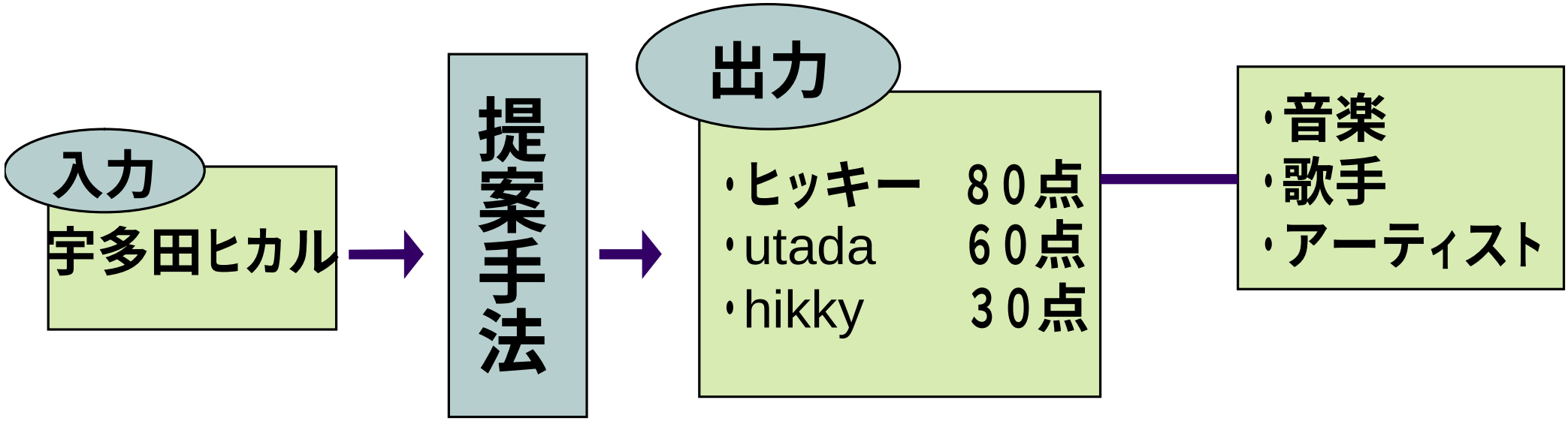


別名 → 再現率の低下



目的

- Webを用いた高精度な別名抽出手法の提案
  - 入力： 人名
  - 出力： (別名, 別名らしさ, キーワード) のリスト
  - 利用可能なもの：
    - 正解データ＝ (名前, 別名リスト) ＊50
    - 検索エンジン



別名？別名らしさ？

- 別名＝人物への参照として用いられる表記

★ 別名らしさ＝どれだけ多くの人が使っているか

~~野球選手~~

~~大人~~

★ 松井昌雄

息子

野球

★★★

ゴジラ

★★ ヤンキース

頼れる4番

野球

★

ゴジちゃん

先行研究

- T. Hokama , et al. , “Extracting mnemonic names of people from the web” , ICADL06.
- **候補の抽出**
  - 「こと<NAME>」で検索→上位５００ページを形態素解析→

    <NAME>で始まる形態素列 (max5)が候補
- **候補のランキング**
  - 「<NAME>」で検索→<NAME>を特徴付ける<pattern>抽出→

    「候補<pattern>」のヒット件数×<pattern>の重みでランキング

  (<pattern>＝形態素列(max3)，重み＝<NAME>との共起の強さ)
- **上位5件を別名とする**

他にもあるのでは？

別名に有効か？

他の評価も考えられるのでは？

提案手法の全体像

1. 別名候補の抽出

2. いろいろな方法で候補を評価

3. 評価の統合

- 堀江貴文

↓

- ライブドア，ホリエモン，ほりえもん
  容疑者，社長，堀江，，，，

↓

- 1位：ほりえもん(30点)　，社長 (30点)
- 2位：ホリエモン(20点)　，ホリエモン (20点)
- 3位：容疑者(15点)　　，ほりえもん (5点)

↓

- 1位：ホリエモン　100点
- 2位：ほりえもん　20点

段階1：別名候補抽出の概要

- 基本的な考え
  - WEBの文書集合のどこかでは，名前の近傍に別名が現れているはず　どう生成するのか？
  - 名前と別名をつなぐパターンを使って，検索すれば別名候補が得られるのでは？
- 例：
  - 「こと蛯原友里」で検索すれば，　どう取るのか？
    ”モデルのエビちゃんこと蛯原友里は．．．”

のような記述を持つスニペットが得られるはず

人名と別名をつなぐパターンの生成

- 正解データから，名前と別名をつなぐパターンを学習

「名前＊別名」
「別名＊名前」
と検索 →

松井 秀喜：つながりキーワード一覧 - 関心空間
ヤンキースの外野手。松井秀喜。通称ゴジラ。実力はご存知のとおり ファンサービスがいいなと感じた 好感がもてる選手。
www.kanshin.com/keyword/801783/connect - 23k - キャッシュ - 関連ページ

パターン			頻度
*	こと	<NAME>	1039
*	”こと	<NAME>	120
*	」こと	<NAME>	109
*	こと、	<NAME>	105
<NAME>	こと	*	87
*	の	<NAME>	60
<NAME>	さん（	*	54
<NAME>	の	*	50

記号除去 →

パターン			頻度
*	こと	<NAME>	1486
<NAME>	さん	*	130
*	の	<NAME>	103
<NAME>	こと	*	95
<NAME>	の	*	93
<NAME>	は	*	72
<NAME>	が	*	67
*	さん	<NAME>	66

パターンの候補取得力の評価

$$F-Measure(p)=\frac{2}{\frac{1}{Precision(p)}+\frac{1}{Recall(p)}}$$

- *Precision(p)*
  - **パターン*p*で正解をとれる確率**
  - (ゴジラ*2 + Godzilla*1) / **全スニペット数**
    = 3 / 6 = 50 %

- *Recall(p)*
  - **パターン*p*で得られる候補の割合**
  - |{ゴジラ,Godzilla}| / **全別名の数**
    = 2 / 3 = 66%

ゴジラ
したという
昨日の
ゴジラ
驚いた
Godzilla

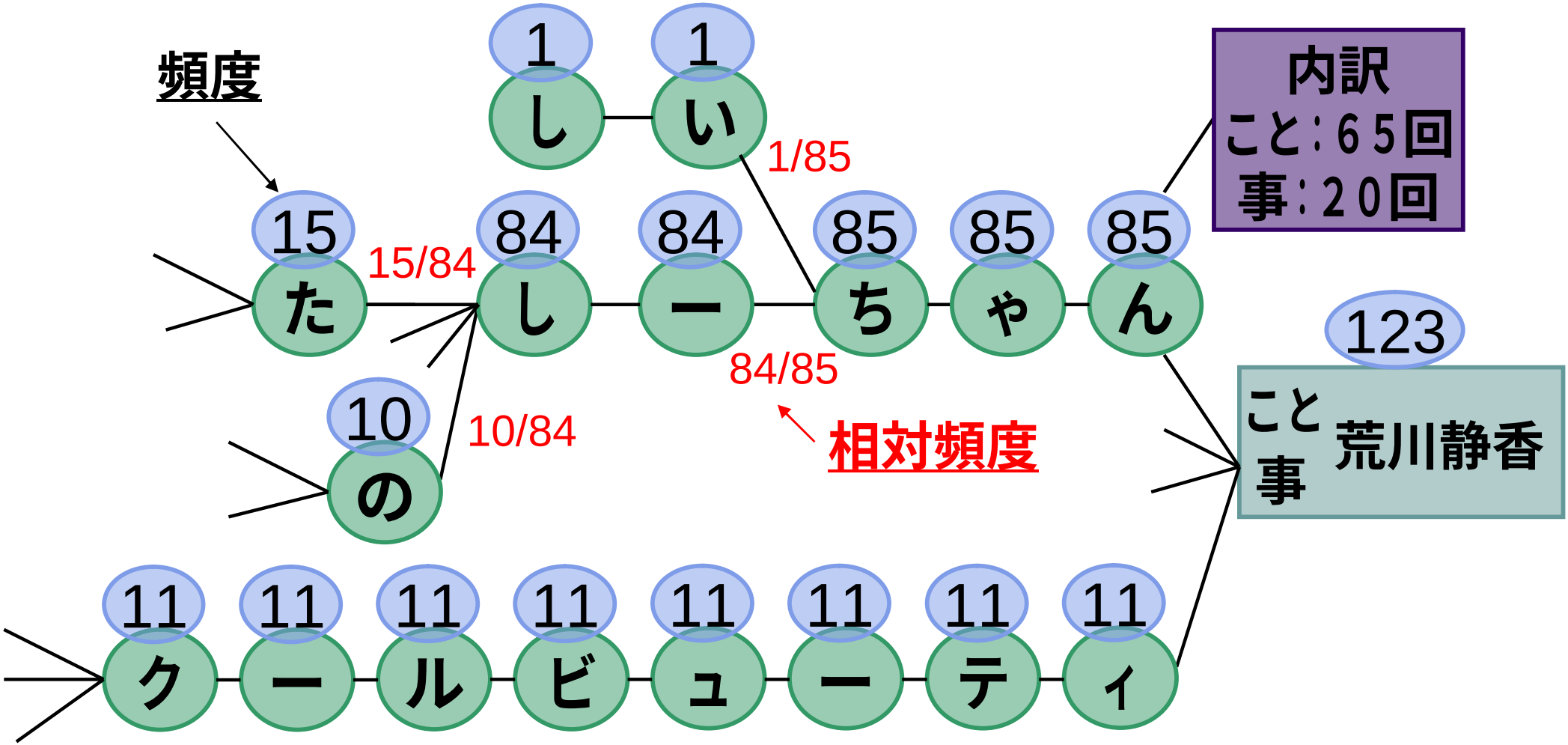
「こと松井秀喜」

パターンの評価結果

順位	パターン			F値	精度	再現率
1	*	こと	<NAME>	0.843	0.762	0.943
2	<NAME>	通称	*	0.551	0.819	0.415
3	<NAME>	こと	*	0.499	0.569	0.444
4	*	事	<NAME>	0.341	0.461	0.270
5	<NAME>	愛称	*	0.284	0.824	0.171
6	<NAME>	以下	*	0.239	0.212	0.276
7	*	ちゃん	<NAME>	0.094	0.071	0.140
8	<NAME>	主演	*	0.075	0.060	0.100
9	<NAME>	は	*	0.071	0.041	0.278
10	<NAME>	社長	*	0.068	0.105	0.050

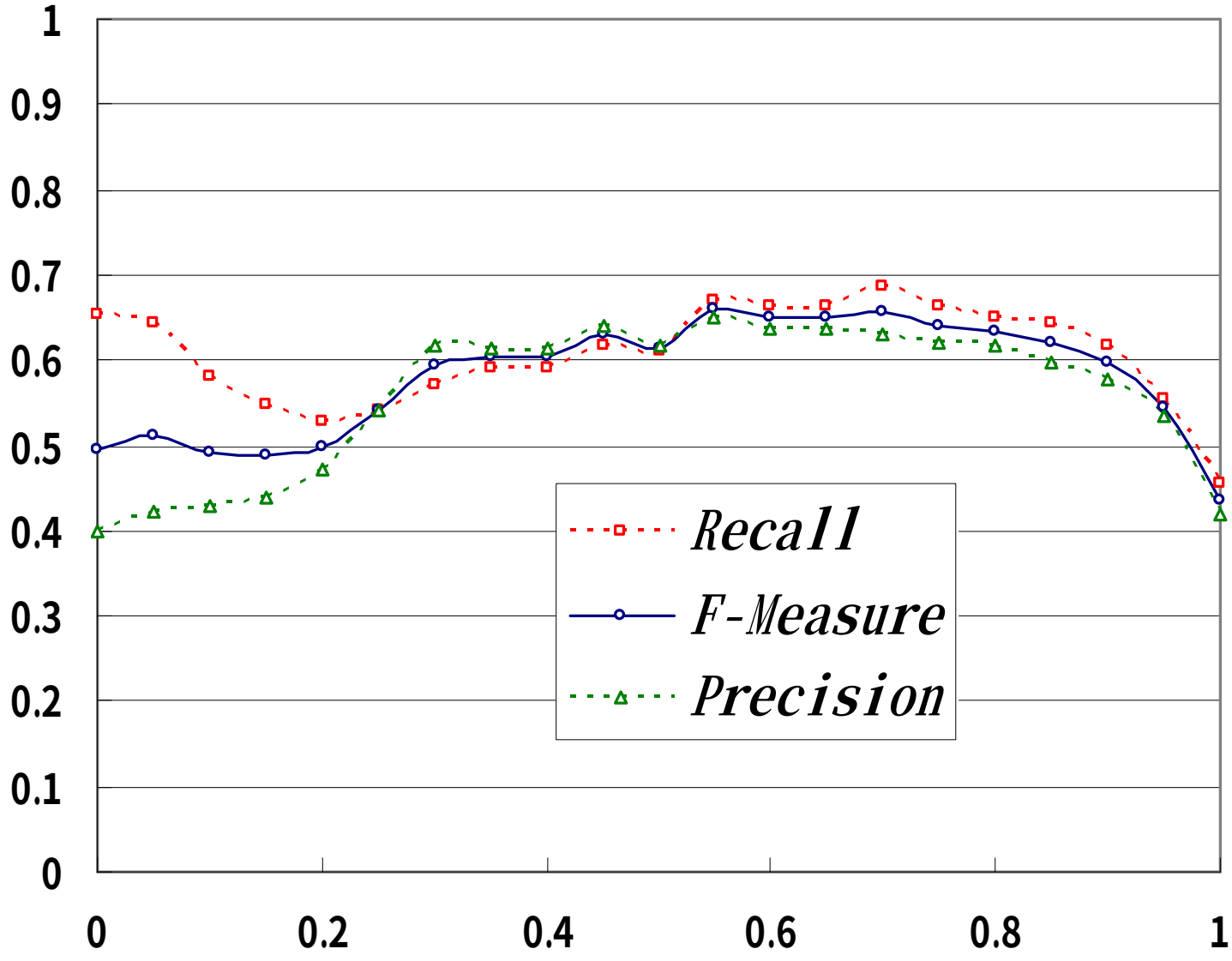
パターンによる候補抽出

- 各パターンで検索してスニペット上位100件を利用



最適な閾値の設定

- ＊こと<NAME>　で調査



段階2：候補の評価の概要

- 本当に別名か？多くの人が使っているか?
  - スニペット内の出現頻度

Freq(ホリエモン) ＞ Freq(ほりえもん) ＞ Freq(の)

  - 名前と候補のdistributional similarity

堀江貴文 ---- {ライブドア，逮捕，買収，IT，…}
ホリエモン---- {ライブドア，オタク,逮捕,成り上がり，…}
社長 ---- {企業，経営，株，年商，…}
Sim(堀江貴文，ホリエモン) ＞ Sim(堀江貴文，社長)

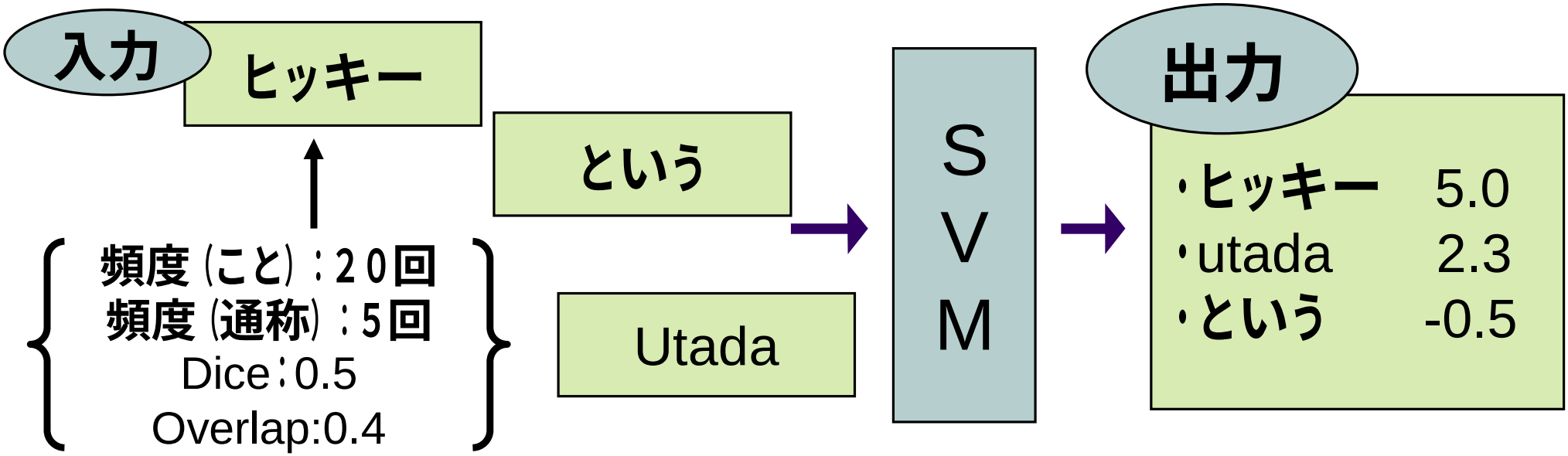
  - 名前と候補のWEB全体での共起性

$$\frac{Hits(\text{堀江貴文}\ and\ \text{ホリエモン})}{Hits(\text{堀江貴文})+Hits(\text{ホリエモン})} > \frac{Hits(\text{堀江貴文}\ and\ \text{社長})}{Hits(\text{堀江貴文})+Hits(\text{社長})}$$

ランキングSVMによる別名候補評価

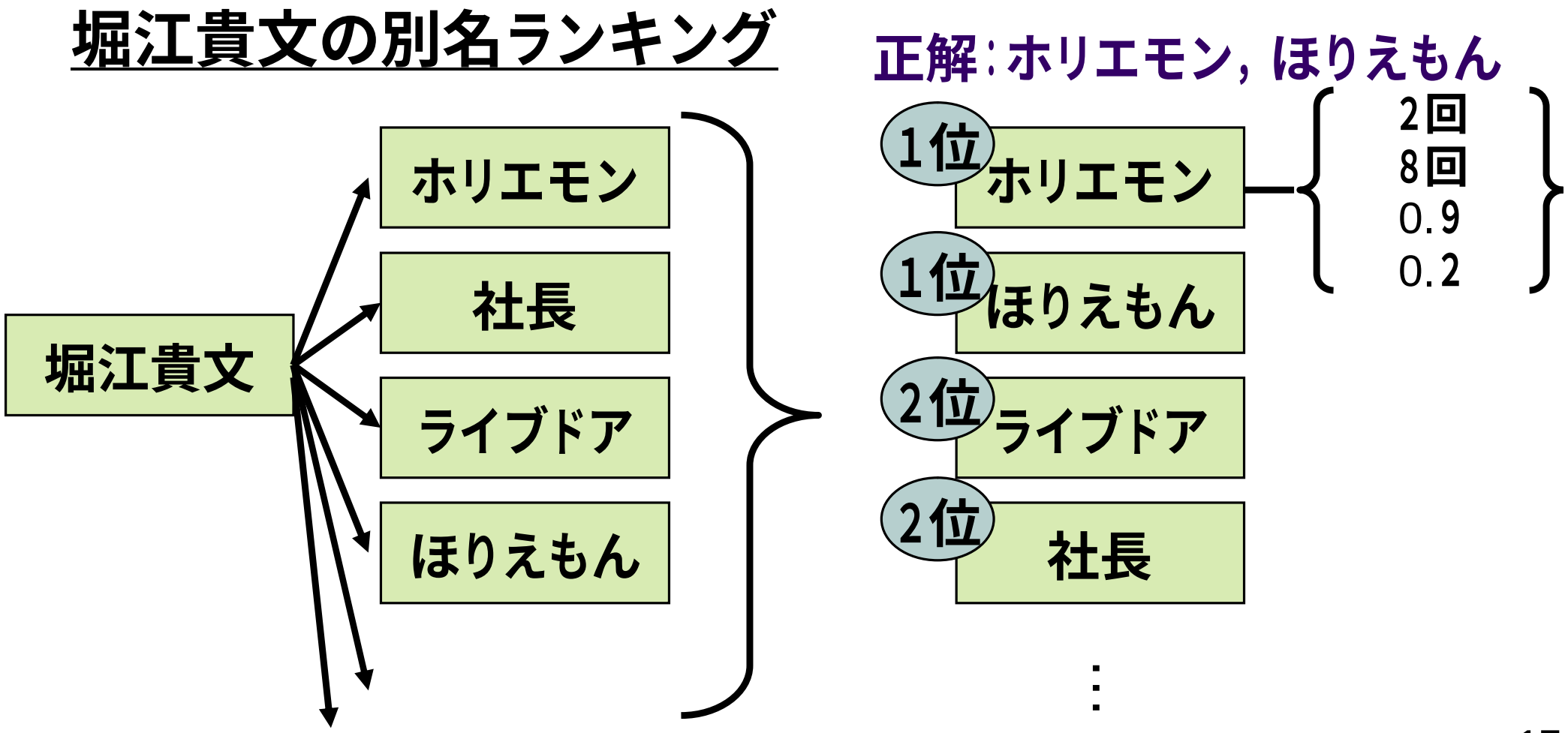
- 入力 → 名前に対する別名候補上位10件
- 候補の表現 → 各評価指標値を並べたベクトル
- 出力 → 別名候補のランキング

<u>宇多田ヒカルの別名が欲しい</u>



ランキングSVMの学習

- 正解データ1件ごとに学習用ランキング生成



ランキングの評価

- ＭＲＲの定義･計算例

$$MRR = \frac{\sum_{Alias \in ExtractedAliases} \frac{1}{Rank(Alias)}}{\sum_{i=1}^{totalAliases} \frac{1}{i}}$$

$$= \frac{\frac{1}{2} + \frac{1}{3}}{\frac{1}{1} + \frac{1}{2}} = 0.556$$

ランク	別名候補
1	の
2	ホリエモン
3	ほりえもん
4	容疑者
5	堀江
6	た

- 各ランキング手法の評価
  - テスト用正解データ50件で平均
  - ランキングＳＶＭが最も良い結果

ランキング手法	*MRR*
SVM(Linear)	0.762
SVM(RBF)	0.729
頻度（こと）	0.702
Dice	0.650
頻度（通称）	0.194

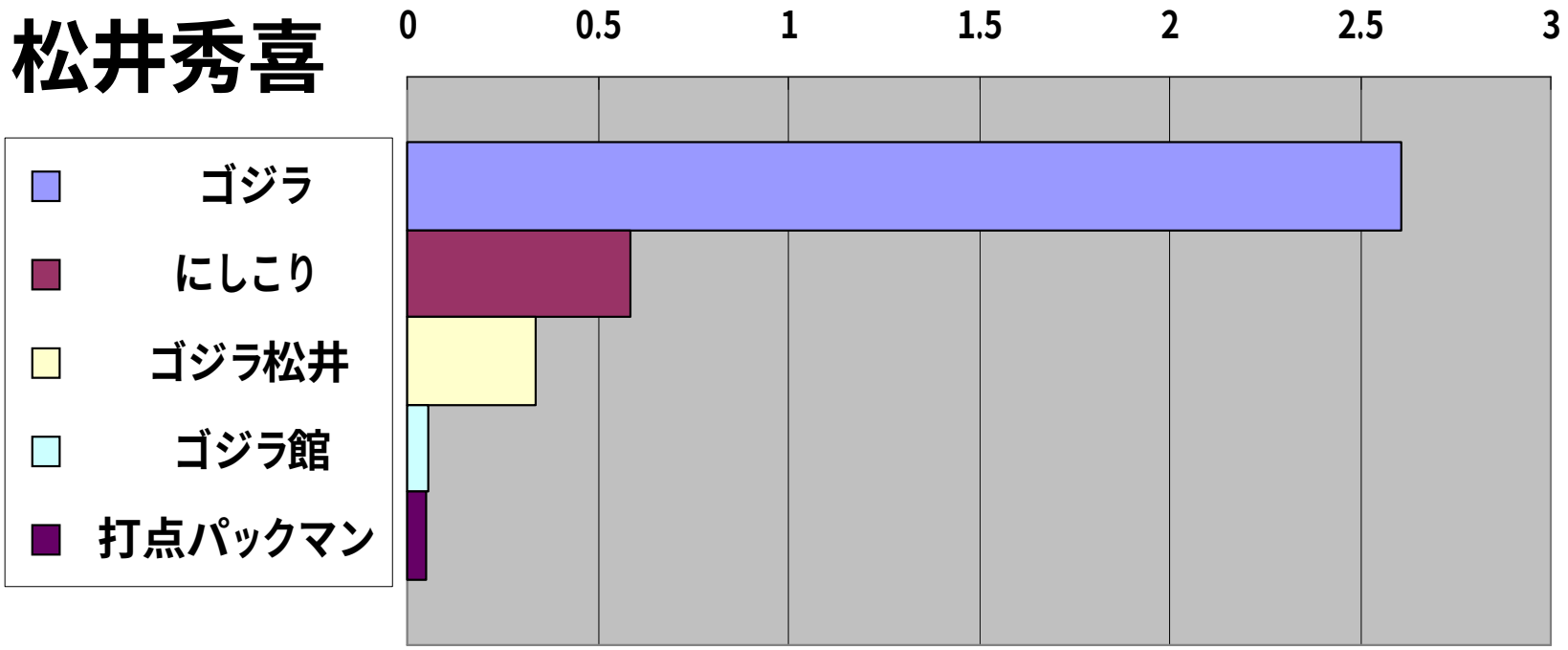
先行研究

・松井

・ゴジラ松井

・ゴジラ

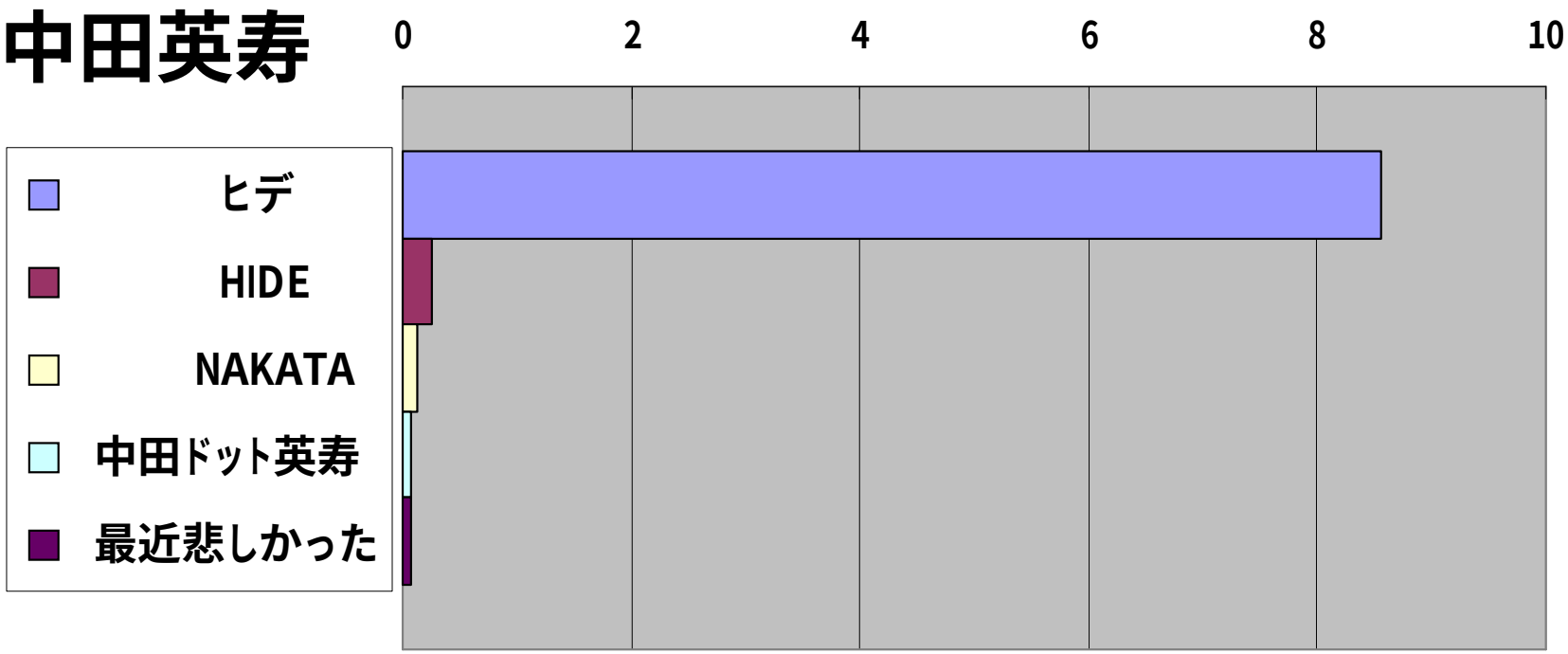
・マツイ



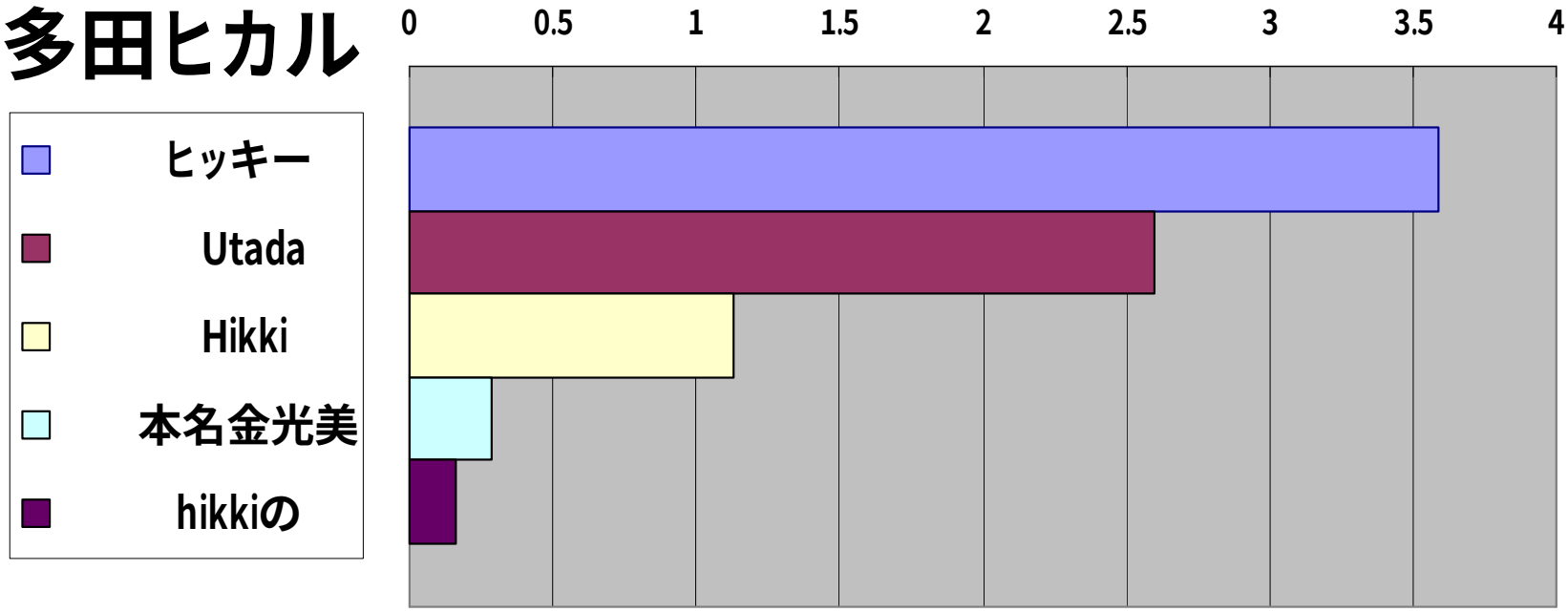
先行研究

・ヒデ

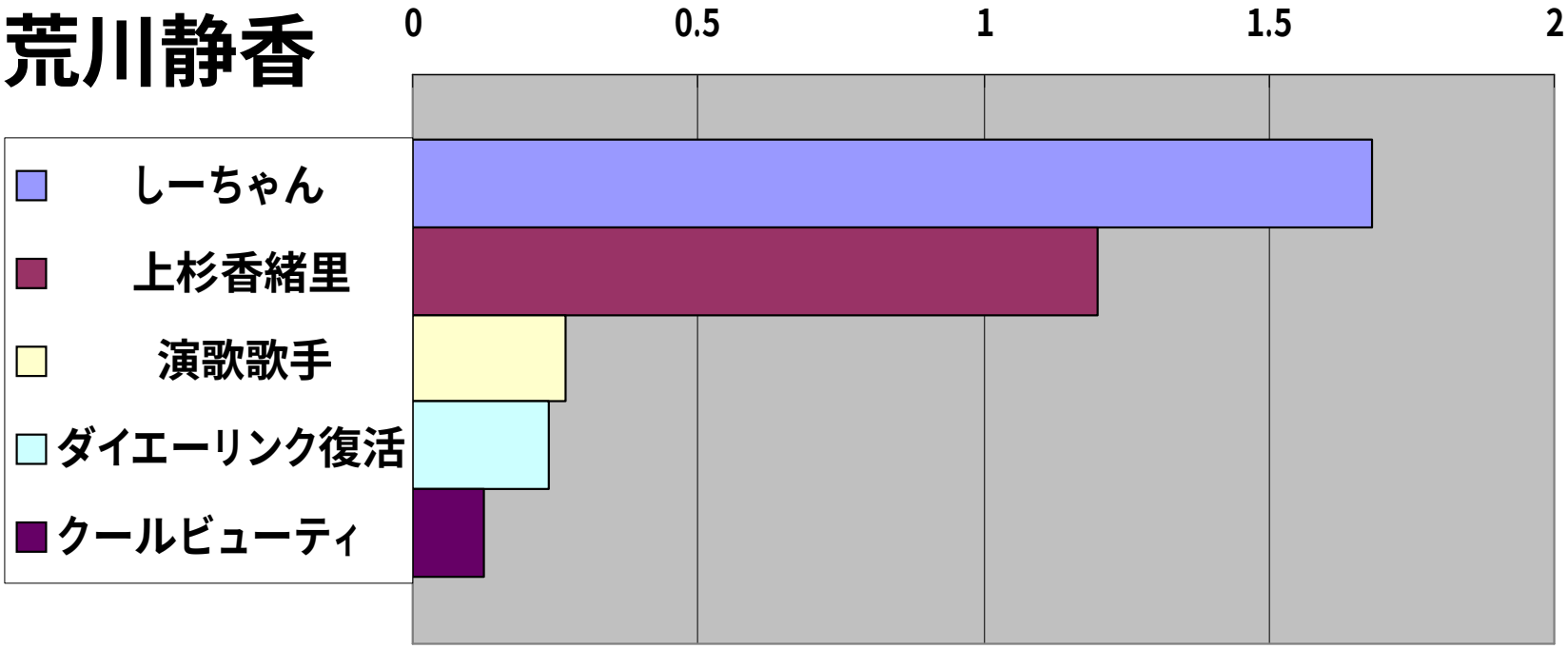
・Hide



宇多田ヒカル



荒川静香



先行研究

- ちゃん
- イナバウアー
- ビューティー
- ーちゃん
- クールビューティー