

Semi-Supervised Clustering に基づく Web 検索結果における人名の曖昧性解消

杉山 一成[†] 奥村 学[‡]

^{†‡} 東京工業大学 精密工学研究所

[†] sugiyama@lr.pi.titech.ac.jp [‡] oku@pi.titech.ac.jp

1 はじめに

World Wide Web (WWW) 上の情報は増え続けているため、ユーザが適切な情報を見つけることは、ますます困難になっている。こうした状況において、Web 検索エンジンは、WWW 上の情報を検索するための有用な手段である。例えば、PageRank [1] や HITS [2] といったアルゴリズムでは、WWW のハイパーリンク構造を利用して、Web ページを重み付けすることにより、高い検索精度を実現している。しかし、検索エンジンは、検索語に応じて長い検索結果のリストを返すので、ユーザが自分の要求に適合する文書を見つけるには、タイトルとスニペットを一つずつ調べていく必要がある。これは、検索語に対する複数の話題が混在している場合には多くの時間を要する作業となる。また、検索語はいくつかの意味を持ち得るが、現在の検索エンジンは、これらの意味に応じて結果を分類することができず、単に結果のリストを返すのみである。例えば、ユーザが、人名である “William Cohen” を検索語として検索エンジン Google に入力した場合、その検索結果には、情報科学の教授、アメリカの政治家、外科医、生物学の教授、歴史家といった人物が、各人物ごとに分類されずに混在している。このような場合、もし検索結果が、検索語の異なる意味に基づいて分類されていたならば、ユーザは自分の要求に適合する文書を容易に見つけることができるであろう。特に、人名は検索語として、検索エンジンにしばしば入力され¹、上述した “William Cohen” の例のように、多くの場合、いくつかの同姓同名の人物を有する。そこで、本論文では、Web の検索結果における人名の曖昧性を解消するための手法を提案する。ほとんどの人物の検索結果は、多くの異なる実体を含み、各実体は一つ、あるいは二つの Web ページから構成される。そこで、semi-supervised clustering を用いて、人名の曖昧性の解消を行う。我々の手法は、semi-supervised clustering において、クラスタの重心の変動を抑える点において新規性がある。

2 関連研究

2.1 統計学習に基づいた手法

人名の曖昧性を解消するための統計学習に基づいた手法は、(1) 教師なし学習による手法、特にクラスタリングに基づいた手法、(2) 教師あり学習による手法、の 2 つに分類することができる。

2.1.1 教師なし学習による手法

クラスタリングに基づいた人名の曖昧性に関する研究について、以下に述べる研究を挙げることができる。

Mann ら [3] は、まず、誕生日、出生地、職業などの人物情報を抽出する。次に、各文書に対して、抽出した人物情報、固有名、文書から計算した TF-IDF スコアからなる特徴ベクトルを作成する。最終的に、この特徴

ベクトルを用いてクラスタリングを行い、人名の曖昧性を解消する。また、Wan ら [4] の研究においても、同様の手法が用いられている。これらの手法では、人物情報を必要とするので、こうした情報を含んでいる文書にしか適用できない。

Pedersen ら [5] は、曖昧性のある名前の文脈に基づいて、文脈ベクトルを作成し、このベクトルを用いてクラスタリングを行う手法を提案している。また、曖昧性のある名前の前後の広い文脈を用いることが有効であることを示している。

Bekkerman ら [6] は、Web の検索結果における人名の曖昧性を解消するために、(1) Web ページのハイパーリンクに基づいた手法、(2) 凝集的クラスタリングを用いた手法、(3) (1) と (2) を組み合わせた手法の 3 つの教師なしの手法を提案している。これらの手法では、曖昧性を解消すべき人物のつながりが既知であることが仮定されている。したがって、事前にこうした情報があれば、提案されている手法は有効であると考えられるが、多くの場合、曖昧性を解消すべき人物のつながりについての知識が与えられているわけではない。

Bollegala ら [7] は、まず、文書集合の凝集的クラスタリングを行い、異なる同姓同名人物を識別するために、得られたクラスタから重要語句を抽出する。この重要語句は、C-value [8] を用いて抽出され、その抽出された語句間の類似度に応じて、文書をクラスタリングする。しかし、C-value を用いると、必要以上に長く、無意味な語句が抽出される傾向にある。したがって、より有用な語句を抽出するために、いくつかのヒューリスティクスを導入し、TF-IDF に基づいた手法よりも正確であると結論付けている。

2.1.2 教師あり学習による手法

Han ら [9] は、naive Bayes や support vector machines を用いて、Web ページ、DBLP の書誌情報における著者の曖昧性を解消する問題を扱っている。提案手法は論文リストに着目しており、共著者の情報が、人名の曖昧性の解消に有効であるという有益な知見を得ている。

Al-Kamha ら [10] は、人名の検索結果における Web ページが同一人物に対応するように、この Web ページをグループ化するための教師あり学習による手法を提案している。この手法では、属性、リンク、検索結果における Web ページの類似度といった三つの側面を考慮することによって、確率を推定するための訓練データを用意する。次に、この訓練データを用いて、各側面について、Web ページの組に対する信頼度行列を作成する。また、Stanford certainty theory [11] を用いて、各側面に対する信頼度行列を結合することで、最終的な信頼度行列を作成する。同じ人物について述べた Web ページは、信頼度行列の値が閾値よりも大きな Web ページの組をマージすることでグループ化される。この手法は、電話番号、メールアドレス、郵便番号といった属性

¹<http://tap.stanford.edu/PeopleSearch.pdf>

を抽出するために多くのヒューリスティクスが導入されている。こうした属性は必ずしも検索された文書に現れるわけではないため、このような情報を含む文書にしか適用できない。これは、2.1.1 節で述べた Mann ら、Wan らの研究と同様の問題を抱えていることになる。

2.2 複数文書間の人名の曖昧性の解消

複数文書間の名前の曖昧性解消は、エンティティやイベントの追跡といった、文書からの知識発見におけるさまざまなタスクに対して必要とされる。この分野における初期の研究において、Wacholder ら [12] は、一つの文書内に共起する名前の知識を用いることによって、固有名を同定し、その曖昧性を解消するための手法を開発している。この手法は、後続する研究において、複数文書の場合に拡張されている。Bagga ら [13] は、まず、各文書における各エンティティに対する共参照のつながりを形成するために、文書内の共参照関係を解決する。次に、各参照のつながりの周囲のテキストを用いて、特徴ベクトルを作成し、クラスタリングを行う。しかし、用いられている特徴は共起語のみであるため、Niu ら [14] は、共起語と自然言語情報の抽出結果の両方を用いた新たな手法を提案している。Gooi ら [15] は、Bagga らの手法を再実装し、大規模コーパスを用いたときには、Bagga らの手法がうまく機能しないことを示した。そこで、*Person-X* コーパスと呼ばれる独自のコーパスを作成し、凝集的クラスタリングを用いて、良い精度が得られるとの結論を得ている。

2.3 外部知識に基づいた人名の曖昧性の解消

人名の曖昧性解消は、二つ以上のあらかじめ決められた選択肢の中から、曖昧性のある語の正しい意味を選択する語義の曖昧性解消の一例とみなすことができる。これらの意味は、WordNet² [16] のようなあらかじめ編集された辞書から導出される。WordNet に関する主な問題は、未知語を扱うことができない点にある。Web の検索結果における人名の曖昧性の解消の場合には、異なる名前の数が、WordNet における未知語の数よりも大きくなり、その数を見積もることはできない。また、語義の曖昧性解消では、上述したように辞書を使用する。一方で、特定の人名に対する同性同名人物に関する情報を、事前に得ることはできない。

外部知識を用いる別の研究において、Bunescu ら [17] は、Wikipedia の構造を用いて、固有名を同定し、曖昧性を解消するための手法を提案している。

3 人名の検索結果に対する Semi-Supervised Clustering

複数の同性同名人物を含む WWW からの人物の検索結果は、次の二つのタイプに分類することができる。

- 検索結果に多くの実体が含まれ、各実体は一つあるいは二つのページから構成される。
- 検索結果にはほとんど実体が含まれないが、各実体は 10 ページ以上から構成される。

WWW における人物の検索結果のほとんどは、(a) に相当する。したがって、もし種文書を導入すれば、人名の曖昧性を解消するためのクラスタリングを、より正確に行うことができると期待される。そこで、本論文では、Web 検索結果における人名の曖昧性を解消するた

めに、semi-supervised clustering を適用する。本研究では、次の 2 つの文書を種文書とし、これらの有効性について検証する。

- Web の検索結果中の実際の人物について述べている Web ページ、
- Wikipedia における各人物の記述。

以下、(1) を “entity seed”，(2) を “wiki seed” と呼ぶことにする。

また、semi-supervised clustering において高い精度を得るには、クラスタの重心に新たなクラスタがマージされるときに、重心の変動を抑えることが重要である。そこで、重心の変動を抑えるために、(1) ユークリッド距離に基づいた手法、(2) マハラノビス距離に基づいた手法を提案する。2.1.1 節で述べたクラスタリングに基づいた手法とは異なり、我々の手法は、semi-supervised clustering における重心の変動を抑える点において新規性がある。

以下、検索結果集合における Web ページ p の特徴ベクトル w^p を式 (1) のように表す。

$$w^p = (w_{t_1}^p, w_{t_2}^p, \dots, w_{t_m}^p) \quad (1)$$

ここで、 m は Web ページ集合中における単語の異なり数であり、 t_k ($k = 1, 2, \dots, m$) は、各単語を表す。また、クラスタの重心ベクトル G を式 (2) のように定義する。

$$G = (g_{t_1}, g_{t_2}, \dots, g_{t_m}) \quad (2)$$

同様に、 m は重心ベクトル中における単語の異なり数であり、 t_k ($k = 1, 2, \dots, m$) は、各単語を表す。

3.1 ユークリッド距離を用いたクラスタ重心の変動の抑制

クラスタの重心ベクトル G と、検索結果集合における Web ページ p の特徴ベクトル w^p 間のユークリッド距離 $D_{euc}(G, w^p)$ は、式 (3) で表される。

$$D_{euc}(G, w^p) = \sqrt{\sum_{k=1}^m (g_{t_k} - w_{t_k}^p)^2} \quad (3)$$

式 (2)、(3) から、重心ベクトルに別のクラスタをマージした後の新たなクラスタの重心ベクトル G^{new} は、式 (4) のように定義される。

$$G^{new} = \frac{G + \frac{w^p}{D_{mhl}(G, w^p)}}{n + 1} \quad (4)$$

ここで、 n は別のクラスタをマージする前のクラスタの重心における文書の数である。

3.2 マハラノビス距離を用いたクラスタ重心の変動の抑制

マハラノビス距離は、データの相関を考慮しているという点で、ユークリッド距離とは異なる。したがって、この距離を用いることで、3.1 節で述べたユークリッド距離を用いるよりも、クラスタの重心の変動をより抑えることができると考えられる。ここで、 Σ を分散共分散行列とすると、クラスタの重心ベクトル G と、検索結果集合における Web ページ p の特徴ベクトル w^p 間のマハラノビス距離 $D_{mhl}(G, w^p)$ は、式 (5) で表される。

$$D_{mhl}(G, w^p) = \sqrt{(w^p - G)^T \Sigma^{-1} (w^p - G)} \quad (5)$$

²<http://wordnet.princeton.edu/>

表 1: 実験データの概要

Personal Name	Number of namesakes	Number of pages about the person
William Cohen	16	98
David Israel	29	65
Bill Mark	9	20
Tom Mitchell	29	76
David Mulford	8	73
Andrew Ng	22	55
Fernando Pereira	11	55
Jim Clark	22	96
Michael Jackson	7	92
Noam Chomsky	1 (linguist and activist)	93
Overall	154	723

式 (2), (5) から, 別のクラスタを重心ベクトルにマージした後の新たなクラスタの重心ベクトル G^{new} は, 式 (6) のように定義される.

$$G^{new} = \frac{G + \frac{w^p}{D_{mhl}(G, w^p)}}{n + 1}, \quad (6)$$

ここで, n は別のクラスタをマージする前のクラスタの重心における文書の数である.

4 実験

4.1 実験データ

本研究では, 先行研究 [6], [7] で用いられている 10 人の人名を実験データとして用いた. 検索エンジン Google に各人名を検索語として入力し, 検索結果として返される上位 100 件の Web ページを収集した. 実験データの概要を表 1 に示す. これらの実験データにおいて, 例えば, “Andrew Ng” は情報科学の教授, 物理学の教授, フットサル選手などの同姓同名人物がある. また, “David Israel” の場合には, ほとんど同じ研究分野に携わる光学の研究者と電子工学の教授が存在する.

4.2 評価尺度

本研究では, クラスタの評価尺度として, “entropy” と “purity” を用いる. entropy は, さまざまなクラスタの文書が各クラスタ内にどれだけ含まれるかを評価する. 一方, purity は, 各クラスタが主に一つのクラスからの文書をどの程度含むかを評価する.

大きさ n_r のクラスタ C_r について, entropy は式 (7) で定義される.

$$E(C_r) = -\frac{1}{\log q} \sum_{i=1}^q \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \quad (7)$$

ここで, q はデータ集合中のクラス数, n_r^i は r 番目のクラスタに割り当てられた i 番目のクラスの文書数である. 全クラスタの entropy はクラスタ数に応じて重み付けされた各クラスタの entropy の和として, 式 (8) で定義される.

$$Entropy = \sum_{r=1}^k \frac{n_r}{n} E(C_r) \quad (8)$$

一般に, entropy の値が小さいほど, 正確なクラスタが生成されたことを示す.

同様に, クラスタ C_r の purity は式 (7) で定義される.

$$P(C_r) = \frac{1}{n_r} \max_i(n_r^i) \quad (9)$$

全クラスタの purity はクラスタ数に応じて重み付けされた各クラスタの purity の和として, 式 (10) で定義される.

$$Purity = \sum_{r=1}^k \frac{n_r}{n} P(C_r) \quad (10)$$

一般に, purity の値が大きいほど, 正確なクラスタが生成されたことを示す.

4.2.1 予備実験

より高いクラスタリングの精度を得るために, 単語の重み付け手法を比較し, クラスタリングのための特徴ベクトル生成に最も有効な方法を決定する. これを確認するために, 我々は, (1) TF, (2) IDF, (3) residual IDF, (4) TF-IDF, (5) x^l -measure, (6) gain, の六つの単語の重み付け手法を用いて特徴ベクトルを作成し, 凝集的クラスタリングを行った. また, 最も類似したクラスタをマージするアルゴリズムとして, 最短距離法, 最長距離法, 群間平均法について比較した.

実験の結果, クラスタリングのための特徴ベクトルを生成するための最も有効な単語重み付け手法は gain であり (entropy: 0.264, purity: 0.703), 最も類似したクラスタをマージするには, 最長距離法が最も効果的なアルゴリズムであることがわかった. gain を用いた場合, 式 (1) で定義される w^p の各要素 $w_{t_k}^p$ は, 式 (11) のように表される.

$$w_{t_k}^p = \frac{df(t_k)}{N} \left(\frac{df(t_k)}{N} - 1 - \log \frac{df(t_k)}{N} \right) \quad (11)$$

ここで, $df(t_k)$ は単語 t_k が出現する Web ページの数を, N は全 Web ページ数を表す.

4.2.2 Semi-supervised Clustering の結果

提案手法の有効性を確認するために, これを重心の変動を抑えない場合と比較する.

表 2, 3 に, それぞれ, entity seed を用いて semi-supervised clustering によって得られた entropy と purity の結果を示す. 同様に, 表 4, 5 に, wiki seed を用いて得られた結果を示す. 4.2 節で述べたように, entropy が小さいほど, また purity が大きいほど, 正確なクラスタが生成されたことを示す.

表 2~5 によれば, entity seed と wiki seed の両方とも, ユークリッド距離, マハラノビス距離を用いてクラスタの重心の変動を抑える方法は, その変動を抑えない方法よりも良い結果が得られたことがわかる. また, entity seed と wiki seed の両方とも, 表 2, 4 に示した entropy の値は, gain で特徴ベクトルを作成し, 最長距離法を用いてクラスタリングを行って得られた entropy の値 (0.264) と比べて改善した. 一方, 表 3, 5 に示した purity の値も, gain で特徴ベクトルを作成し, 最長距離法を用いてクラスタリングを行って得られた purity の値 (0.703) と比べて改善した.

以上の結果から, 次の知見が得られる. 種文書を導入したことによって, 適合しない検索結果が, その実体の検索結果に含まれなくなったことにより, entropy が改善された. また, 0.80 に近い purity の値が得られていることから, 種文書として entity seed, wiki seed を使用することは適切であると考えられる. さらに, マハラノビス距離によってクラスタの重心の変動を抑える手法において, entity seed, wiki seed を用いて得られた purity の値の差は小さい. したがって, Wikipedia における各人物の記述が継続的に更新されれば, 種文書の情報源としての信頼性は改善されるものと考えられる.

5 おわりに

本論文では, Web 検索結果における人名の曖昧性を解消するために, semi-supervised clustering に基づいた手法を提案した. クラスタリングのための特徴ベクトルを生成するには, gain がこの目的に適合する最も有効な単語の重み付け法であることがわかった. また, 人物の検索結果のほとんどは, 検索結果に多くの実体

表 2: entity seed を用いた entropy の結果

Personal name	Uncontrolled	Controlled by Euclidean distance	Controlled by Mahalanobis distance
William Cohen	0.100	0.099	0.098
David Israel	0.151	0.148	0.146
Bill Mark	0.125	0.082	0.080
Tom Mitchell	0.219	0.178	0.176
David Mulford	0.239	0.219	0.217
Andrew Ng	0.223	0.152	0.150
Fernando Pereira	0.294	0.232	0.231
Jim Clark	0.166	0.137	0.134
Michael Jackson	0.121	0.102	0.101
Noam Chomsky	0.180	0.156	0.154
Average	0.182	0.150	0.149

表 3: entity seed を用いた purity の結果

Personal name	Uncontrolled	Controlled by Euclidean distance	Controlled by Mahalanobis distance
William Cohen	0.864	0.865	0.868
David Israel	0.669	0.672	0.673
Bill Mark	0.894	0.905	0.905
Tom Mitchell	0.701	0.720	0.734
David Mulford	0.743	0.746	0.748
Andrew Ng	0.694	0.729	0.732
Fernando Pereira	0.675	0.748	0.750
Jim Clark	0.766	0.800	0.802
Michael Jackson	0.912	0.923	0.925
Noam Chomsky	0.913	0.916	0.918
Average	0.783	0.802	0.805

を含み、各実体は1つあるいは2つのWebページから構成されるため、semi-supervised clusteringを導入した。我々の手法は、semi-supervised clusteringにおいて、高い精度を達成するために、クラスタの重心の変動を抑える点に新規性がある。提案する semi-supervised clustering では、ユークリッド距離、マハラノビス距離を用いて、クラスタの重心の変動を抑える工夫をした。この工夫により、Web 検索結果における人名の曖昧性解消をより正確に行うことができると結論付けられる。特に、Wikipedia における各人物の記述は、semi-supervised clustering の種文書として、有望な情報源であると考えられる。

本論文では、検索結果集合における Web ページ中の各単語の重み付けを、gain を用いて行った。しかし、gain によって求められる値が、ある人物を特徴付ける単語のスコアに必ずしも高い値を割り当てるとは限らない。したがって、より正確な人名の曖昧性解消を行なうために、検索結果の Web ページ中の単語に、きめ細かい重みを割り当てる手法について、今後研究を進めていく予定である。

参考文献

- [1] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report SIDL-WP-1999-0120, Stanford Digital Library Technologies Project, 1998.
- [2] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM (JACM)*, Vol. 46, No. 5, pp. pages 604–632, 1999.
- [3] G. S. Mann and D. Yarowsky. Unsupervised Personal Name Disambiguation. In *Proc. of the 7th Conference on Natural Language Learning (CoNLL-2003)*, pp. 33–40, 2003.
- [4] X. Wan, J. Gao, M. Li, and B. Ding. Person Resolution in Person Search Results: WebHawk. In *Proc. of the 14th International Conference on Information and Knowledge Management (CIKM '05)*, pp. 163–170, 2005.
- [5] T. Pedersen, A. Purandare, and A. Kulkarni. Name Discrimination by Clustering Similar Contexts. In *Proc. of the 6th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing2005)*, pp. 226–237, 2005.
- [6] R. Bekkerman and A. McCallum. Disambiguating Web Appearances of People in a Social Network. In *Proc. of*

表 4: wiki seed を用いた entropy の結果

Personal name	Uncontrolled	Controlled by Euclidean distance	Controlled by Mahalanobis distance
William Cohen	0.126	0.120	0.121
David Israel	0.228	0.164	0.158
Bill Mark	0.158	0.129	0.127
Tom Mitchell	0.278	0.120	0.104
David Mulford	0.232	0.159	0.155
Andrew Ng	0.139	0.141	0.139
Fernando Pereira	0.270	0.147	0.145
Jim Clark	0.129	0.129	0.119
Michael Jackson	0.112	0.082	0.081
Noam Chomsky	0.164	0.104	0.101
Average	0.184	0.130	0.123

表 5: wiki seed を用いた purity の結果

Personal name	Uncontrolled	Controlled by Euclidean distance	Controlled by Mahalanobis distance
William Cohen	0.870	0.871	0.873
David Israel	0.590	0.620	0.635
Bill Mark	0.642	0.692	0.717
Tom Mitchell	0.660	0.640	0.645
David Mulford	0.740	0.760	0.770
Andrew Ng	0.690	0.750	0.780
Fernando Pereira	0.730	0.810	0.850
Jim Clark	0.730	0.790	0.793
Michael Jackson	0.900	0.900	0.905
Noam Chomsky	0.830	0.850	0.860
Average	0.738	0.768	0.783

the 14th International World Wide Web Conference (WWW2005), pp. 463–470, 2005.

- [7] D. Bollegala, Y. Matsuo, and M. Ishizuka. Extracting Key Phrases to Disambiguate Personal Names on the Web. In *Proc. of the 7th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing2006)*, pp. 223–234, 2006.
- [8] K. Frantzi and S. Ananiadou and H. Mima. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *International Journal on Digital Libraries*, Vol. 3, No. 2, pp. 107–145, 2000.
- [9] H. Han, L. Giles, H. Zha, C. Li, and K. Tsioutsoulis. Two Supervised Learning Approaches for Name Disambiguation in Author Citations. In *Proc. of ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)*, pp. 296–305, 2004.
- [10] R. Al-Kamha and D. W. Embley. Grouping Search-Engine Returned Citations for Person-Name Queries. In *Proc. of the 6th ACM International Workshop on Web Information and Data Management (WIDM'04)*, pp. 96–103, 2004.
- [11] G. Luger and W. Stubblefield. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Addison Wesley Longman, 1997.
- [12] N. Wacholder, Y. Ravin, and M. Choi. Disambiguation of Proper Names in Text. In *Proc. of the 5th Conference on Applied Natural Language Processing (ANLP 1997)*, pp. 202–208, 1997.
- [13] A. Bagga and B. Baldwin. Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In *Proc. of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL '98)*, pp. 79–85, 1998.
- [14] C. Niu, W. Li, and R. K. Srihari. Weakly Supervised Learning for Cross-document Person Name Disambiguation Supported by Information Extraction. In *Proc. of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004)*, pp. 597–604, 2004.
- [15] C. H. Gooi and J. Allan. Cross-Document Coreference on a Large Scale Corpus. In *Proc. of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, pp. 226–237, 2005.
- [16] G. A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, Vol. 38, pp. 39–41, 1995.
- [17] R. Bunescu and M. Pasca. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pp. 9–16, 2006.