半教師有り学習に基づくWeb上の人物クラスタリングシステム

池田 雅紀† 佐藤 一誠† 吉田 稔† 中川 裕志‡

†東京大学大学院情報理工学系研究科〒113-0033 東京都文京区本郷7-3-1
‡東京大学情報基盤センター

1 はじめに

Web検索において、人物検索の重要度が増すに従って，人物の検索に関する問題として人物の同姓同名問題の解消が求められている．人物の同姓同名問題とは，Web検索において検索対象者と同姓同名の人物の存在によって検索結果から目的の人物のページを発見することが困難になるという問題である．特に困難な場合としては以下の場合が考えられる．第一に，検索対象者と同姓同名の有名人が存在する場合 (例：米国前大統領 "George Bush") である．第二に，検索対象者の名前が多くの同姓同名の人物を持つ場合 (例："田中太郎" や "John Smith") である．このように，同姓同名問題は言語を問わず問題となっている．この問題の解決のため，同一人物ごとに分類し，表示する方法が提案されている．人物を特定する素性は個人差が大きいため，この問題の解決には人物の分類に教師有り学習を用いることは難しく，クラスタリングによる分類が用いられている．同姓同名の人物のクラスタリングには文書中の人物に関わる素性を発見することが重要であり，人名，地名，組織名といった固有表現がクラスタリングにおいて有効であると先行研究 [1] によって示されている．

本稿では，既存のクラスタリングで得られたクラスタから抽出した素性を用いて，同姓同名クラスタリングを行うことを提案する．本手法では，ブートストラップと呼ばれる半教師有り学習の手法を用いることで，抽出した素性を用いて，第一段階のクラスタを拡張する．また，提案手法に対して評価を行い，同姓同名クラスタリングの性能が向上することを確認した．

我々は本手法をWeb人物クラスタリングシステムに組み込み，運用している．本発表では，このシステムについて紹介する．

Person Name Disambiguation System by Semi-supervised Learning

Masaki IKEDA†, Issei SATO†, Minoru YOSHIDA†Hiroshi Nakagawa‡
†Graduate School of Information Science and Technology, The University of Tokyo. Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033 Japan
‡Information Technology Center, The University of Tokyo
{ikeda,sato,mino}@r.dl.itc.u-tokyo.ac.jp, nakagawa@dl.itc.u-tokyo.ac.jp

2 半教師有り学習によるクラスタの拡張

第一段階において，Precisionが高いクラスタの集合が得られたと仮定し，半教師有り学習を用いたクラスタリングについて述べる．文書集合を文書ベクトル $\mathbf{d}$ で表し，これと対応する素性ベクトル $\mathbf{f}$ を考える．このとき，$\mathbf{d}$ と $\mathbf{f}$ との関係は共起行列 P によって表されるとする．次に，任意のクラスタ C に対して，$\mathbf{d}$ が C に属する帰属度を $\mathbf{r}_{d,C}^{(i)}(i=0,\cdots,n)$，$\mathbf{f}$ が C に属する帰属度を $\mathbf{r}_{f,C}^{(i)}(i=0,\cdots,i)$ とする．この時，$\mathrm{r}_{d,C}^{(i)}, \mathrm{r}_{f,C}^{(i)}$ について式 (1),(2) の関係が成り立つ．

$$\mathbf{r}_{f,C}^{(i)} = \mathrm{P}\mathbf{r}_{d,C}^{(i)} \tag{1}$$

$$\mathbf{r}_{d,C}^{(i+1)} = \mathrm{P}^{\mathrm{T}}\mathbf{r}_{f,C}^{(i)} \tag{2}$$

共起行列 P について説明する．Komachi[2] を元に自己相互情報量 pmi を使い，P を式 (3) のように定義する．

$$\mathrm{P}_{i,j} = \begin{cases} \frac{1}{|D||F|}\frac{pmi(d_i,f_j)}{\max pmi} & pmi(d_i,f_j) \geq 0 \\ 0 & pmi(d_i,f_j) < 0 \end{cases} \tag{3}$$

と表される．式 (1),(2) を反復することで，最終的な文書帰属度 $\mathbf{r}_{d,C}^{(n)}$ を得ることができる．

本研究では，半教師有り学習をクラスタ拡張に対して用いるため，複数クラスタに対して適用する手法について提案する．まず，各クラスタ C について，初期の文書帰属度のベクトル $\mathbf{r}_{d,C}^{(0)}$ を，文書がクラスタに所属している場合は1，所属していない場合は0として決める．次に，式 (1),(2) に従って，各クラスタへの帰属度を計算する．各文書 d は帰属度が最大となるクラスタに属する．

3 評価

日本語のWeb上の同姓同名人物クラスタリングにおける，半教師有り学習の評価を行う．実験データは

表 1: クラスタ拡張の評価実験

Topic	BEP	BER	F_B
ONE IN ONE BASELINE	1.00	0.25	0.37
ALL IN ONE BASELINE	0.36	1.00	0.47
1-st	0.84	0.67	0.73
2-nd , Unigram, 反復 1 回	0.80	0.75	0.76
2-nd , Unigram, 反復 2 回	0.66	0.81	0.70
2-nd , Bigram, 反復 1 回	0.84	0.68	0.73
2-nd , Bigram, 反復 2 回	0.83	0.71	0.75

Yahoo API [1]を用いてあらかじめ取得した人物の文書集合 (38 人，文書数最高 193) を用いる．評価は全体を 5 分割した交差検定で行い，Amigó ら [3] の評価手法を用いた．

評価実験として，クラスタ拡張による性能の向上を確認する．対象に本文から固有表現 (人名，地名，組織名)，重要語を抽出し，類似度計算・階層併合クラスタリングを行う．その後，階層併合クラスタリングの結果を用いて，半教師有り学習に基づくクラスタ拡張を行う．素性として，本文の形態素解析から得られた N-gram(N=1,2) を用いる．反復回数を 1,2 回として，実験を行った．

実験結果は表 1 に示した．BEP は Precision，BER は Recall に相当する指標であり，F_B はこれらの F-measure である．評価は F_B によって行う．ONE IN ONE BASELINE は 1 文書 1 クラスタとした場合，ALL IN ONE BASELINE は全文書 1 クラスタとした場合の結果である．一段階目の階層クラスタを行った結果は 1-st に示した．二段階目のクラスタ拡張を行った結果は 2-nd に示し，用いた素性と反復回数を示した．この結果から，2 段階目において，クラスタを拡張することによって，性能が向上していることが確認できる．Unigram を素性とした場合は反復回数が 1 回，Bigram を素性とした場合は反復回数が 2 回の時点で最も良い値となった．

4 Web 検索における同姓同名の曖昧性解消システム Nayose

我々が開発している，Web 検索における同姓同名の曖昧性解消システム Nayose(図 1) について説明する．(http://ianua7.r.dl.itc.u-tokyo.ac.jp:8080/nayose/servlet/Nayose)

[1]http://search.yahooapis.jp/WebSearchService/V1/webSearch

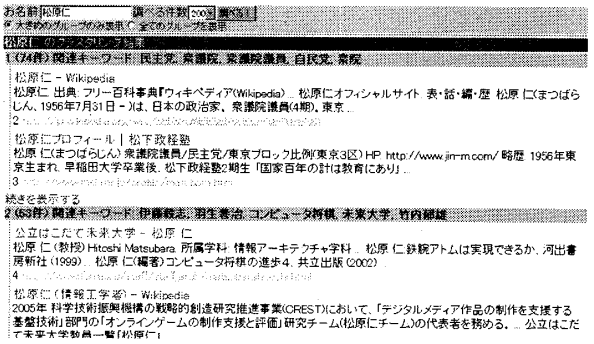

図 1: Nayose デモ

本システムは Yahoo API を用いて取得したスニペット，または，対象 Web ページ本文から固有表現 (人名，地名，組織名)，重要語を抽出し，類似度計算・階層併合クラスタリングを行い，クラスタごとに分類した結果を表示する．本システムでは，本文の形態素解析から得られた N-gram を用いて，半教師有り学習による拡張を行っている．

5 まとめ

本稿では，半教師有り学習の手法を人物のクラスタリングに応用することで，得られたクラスタを拡張する手法を提案し，評価実験によって $F_B = 0.73$ から $F_B = 0.76$ に性能向上することを確認した．また，このクラスタ拡張手法を組み込んだ Web 人物検索システムである Nayose の紹介を行った．

参考文献

[1] E. Elmacioglu, Y. Tan, S. Yan, M. Kan and D. Lee: "PSNUS: Web People Name Disambiguation by Simple Clustering with Rich Features", The SemEval-2007, pp. 268–271 (2007).

[2] M. Komachi, T. Kudo, M. Shimbo and Y. Matsumoto: "Graph-based Analysis of Semantic Drift in Espresso-like Bootstrapping Algorithms", Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pp. 1010–1019 (2008).

[3] E. Amigó, J. Gonzalo, J. Artiles and F. Verdejo: "A comparison of extrinsic clustering evaluation metrics based on formal constraints", Information Retrieval, pp. 1–26 (2008).