異なる配布元からの類似内容文書の発見

谷村正剛† 田中(石井)久美子‡ 中川裕志†

†東京大学 情報基盤センター

‡東京大学 情報学環

1 はじめに

文書の電子化が進むにつれ、新規な文書や語といった、未知のデータの分類に対するニーズが増している。これらは自然言語処理においてはクラスタリングの問題として位置つけることができる。

クラスタリングアルゴリズムとしては多変量解析の分野の最短距離法から始まり、近年では統計的な手法が提案されている [1,2,7]。しかし、クラスタリングの問題として、Manning らは現実には近似解法を導入した上で初期パラメータに依存した局所解しか得ることができないことを指摘している [7]。

クラスタリングの問題自体を見直してみると、その中で扱われている問題はさまざまである。とりわけ、われわれが注目したのはもともと2種類に分類されているデータを別の観点から再分類する問題である。具体的には

- 別の言語で書かれた2つの文書群から似た内容の文書対を抽出する
- 同一の言語で別に作成された文書群から類似の内容をもつ文書対を抽出する

などさまざまな問題が挙げられる。この問題設定は、もともと2種類に分類されているから、その中のデータの対応付けの問題としても捉えることができる。つまり、クラスタリングと対応付けの中間に位置する問題なのである。

文書をもともとの分類にしたがって2部グラフによりモデル化すると、対応付け問題はマッチングを求めることにより解ける。ここに、マッチングとはどの文書も高々1つの文書と対応づけられるような文書対の集合である。マッチングは、Hungarian Method[3,6] などに代表されるように決定論的なアルゴリズムを用いて最適解を求めることができる。類似の関連研究として、Dhillon はこのメリットに着目し、特殊なクラスタリング問題を2部グラフにより解く手法を提案している [4]。つまり、分類問題のうち、特に対応付けとしてもみなすことができる一部の問題に関しては、敢えて解くことが難しいクラスタリングとして位置つけるよりも、扱いやすい対応付けの問題として定式化して解くべきであると考えられる。

このような観点から、われわれはクラスタリング問題のうち、もともと2つに分類されているものを再分類する問題について2部グラフを用いた対応付け問題として扱う方法の枠組を示す。その上で、この枠組を次の実問題に適用した結果を報告する。すなわち、読売新聞および朝日新聞が同じ日に発信した記事の間で、同一内容の記事を対応づける問題を解く。以下、2節にて2部グラフによる文書のモデル化および対応付けを得る手法を説明する。3節では WWW 上の新聞記事を対象としたシステムの実装について述べる。4節ではこのシステムを用いた実験について説明する。最後に5節にてまとめる。

2 モデル化および手法

提案手法におけるモデル化を、図1を用いて説明する。図1の下半分は、もともと2組に分かれる6つの文書を表す。例えば、左側が読売、右側が朝日の各記事となる。文書 A および文書 B が同一内容であるとしたとき、それらを左右の文書の間で対応づける問題を考える。

まず、対応付けの対象とする文書を、図1上半分のように2部グラフのノードとみなす。左側の文書は左側のノード $s_i \subseteq S$、右側の文書は右側のノードを $t_j \subseteq T$ 用いて表す。この枠組においては、左側の任意の文書が右側の任意の文書と対応する可能性がある。このため、すべてのノードの間にエッジを張る。次に、エッジの両端にある文書間の類似度を表す重み $w(i,j)$ を計算し、エッジに与える。$w(i,j)$ の計算方法は、3.2節にて述べる。

以上のような2部グラフから実際に内容が似ている文書を含むエッジを得るため、Hungarian Method [3,6] を用いてマッチングを求める。Hungarian Method [3,6] は、$w(i,j) > 0$ となる重みつき2部グラフにおいて、エッジの重みの総和が最大となるマッチングを求める決定論的な手法である。このようなマッチングは最大重みマッチングと呼ばれる。形式的には、エッジの重みの総和 W_{match} は式 (1) のように表される。

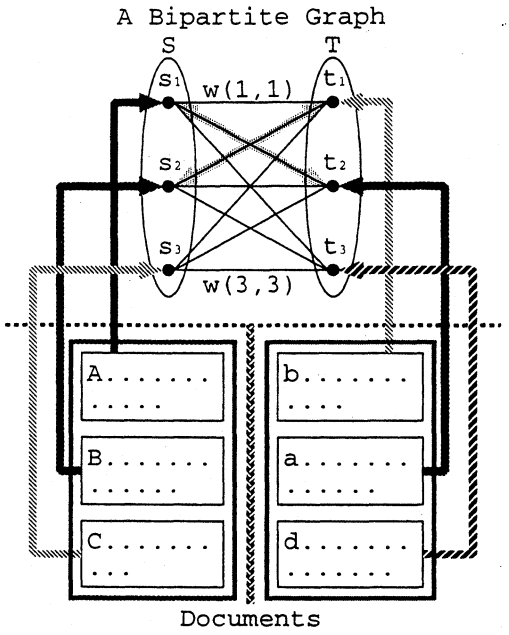


図 1: Example of a Bipartite Graph

$$W_{match} = \sum_{(i,j)\in M} w(i,j) \qquad (1)$$

ここに、M はマッチングに含まれる文書対の集合である。重みつき 2 部グラフについては、式 (2) に示す最大重みマッチングの存在が保証されている [3,6]。

$$\hat{M} = \arg\max_{M} W_{match} \qquad (2)$$

Hungarian Method は、$\hat{M}$ を $O\left(|S|^3 + |T|^3\right)$ の計算量にて求める。

3 実装例

3.1 概要

我々の手法を用い、異なる WWW サイトから配信された同一内容の新聞記事を対応付け、表示するシステムを試作した (図 3)。我々のシステムは、ユーザの入力を受けて駆動する。処理の流れを以下に示す。

1. ユーザが読みたい記事の日付および発信元を図 2(a) のようにシステムに指定する。
2. HTTP クライアントがユーザが指定した日付の記事を WWW から取得する。
3. 取得した各記事について、3.2 節で述べる特徴ベクトルを生成する。

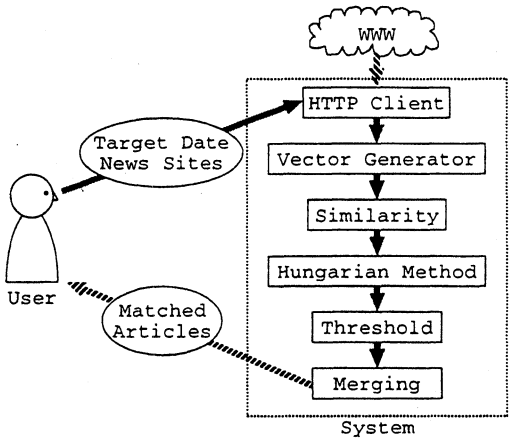


図 3: Our System

4. エッジの重みを求めるため、3.2 節で述べる記事間の特徴ベクトルの類似度を計算する。
5. Hungarian Method を用いて重みつき 2 部グラフの最大重みマッチングを求める。
6. 得たマッチングには類似度が低いものが含まれるため、類似度が 4.3 節で述べる閾値を上回るもののみを対応づけされた記事対として抽出する。
7. 対応づけされなかった記事について、3.3 節で述べる併合を行う。
8. 抽出した記事対について、片方の発信元からの記事のみを図 2(b) のように表示する。必要ならば、記事のタイトルにあるリンクをたどることにより発信元の記事を読むことができる。

我々のシステムは、例えばニュースの読者が複数の新聞社を渡り歩きながら読む際、同じ内容の記事を排除するなどの応用に向けたものである。以下、システムの詳細を述べる。

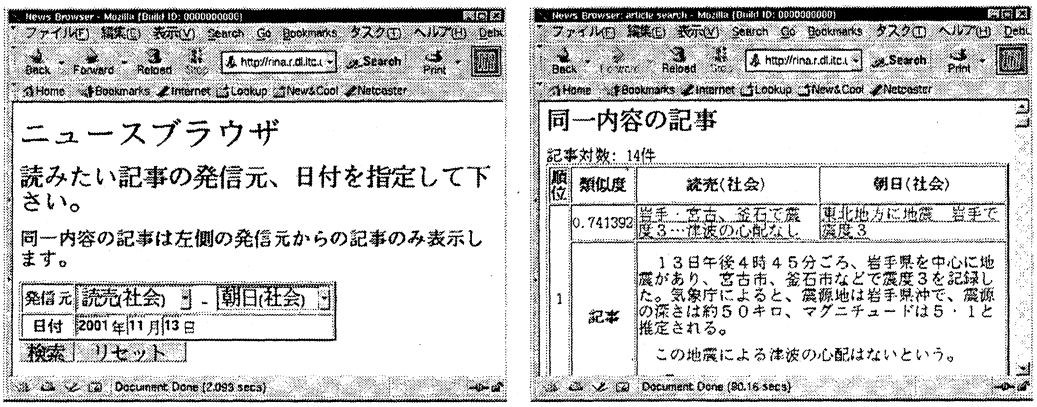
3.2 特徴ベクトルおよび類似度

記事の特徴ベクトルの要素としては様々な方法を試してみた。結果の良さや、新しいパラメータを導入する必要がないことから、以下の 2 種類の値を用いた。

手法 1. 名詞および未知語の TF
手法 2. 名詞および未知語の TFIDF

記事からの単語抽出には、ChaSen 2.02 [8] を用いた。試作したシステムでは、手法 2 を用いた。

記事の長さに対する依存性を避けるため、記事間の類似度には cosine 類似度を用いた。



(a) Date and Sources　　(b) Matched Articles

図 2: Screen Shots of Our System

3.3 併合

我々の手法では、記事に対して 1 対 1 の対応付けをとる。しかし、実験に用いた記事では、同一の新聞社が同じ内容の記事を複数回発信することがしばしばある。このため、閾値以上の記事対を抽出した後、対応づけられなかった記事について、対応付けられた記事の中で最も類似度の高いものを選ぶ。そして、これらの類似記事群はマッチングのとれた記事を代表とする 1 つの記事として扱う。これを併合処理と呼ぶ。

4 実験

我々の手法を評価するため、3 節にて示したシステムを用いて実験を行った。

4.1 新聞記事データ

実験データとして、11 月 7 日から 12 月 6 日までの間に配信された、WWW 版社会面記事 30 日分を用いた。1 日当たりの平均記事数を表 1 に示す。

表 1: Average Number of News Articles per Day

発信元	平均記事数
読売	47.7
朝日	26.1

表 1 に示した記事に対し、人手により対応付けの正解を作成した。正解は 10 日分をひとまとまりとして 3 等分した記事を 3 組の判定者、計 6 名で分担して作成した。正解の基準は NTCIR2 における正解作成手法 [5] に習い、以下のように定めた。

- 2 名が同じ記事を対応づける。
- 2 人が異なる結果を出した場合は、話し合って決める。その結果を元に、我々が最終的な判断を下す。
- 正解はマッチングとなるように作成する。
- 記事の同一性の判断は、記事が伝える事実のみに基づいて行う。記者の主観に基づく表現は判断には用いない。
- 記事が複数の事実を伝えている場合は、最も強く主張されていることを優先する。

4.2 類似度の分布

最初に、マッチングの有効性を確かめるため、マッチングにより得た記事対で類似度がどのように分布しているのかについて調べた。各特徴ベクトル毎に求めた類似度の分布を図 4 に示す。

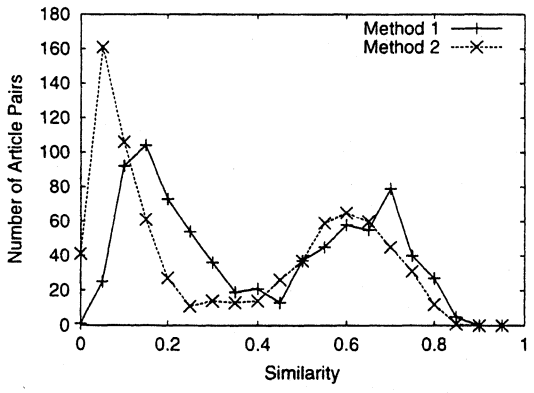


図 4: Distribution of Similarity

特徴ベクトルの種類にも依るが、類似度0.35-0.45を谷として分布がはっきりと分かれている。類似度が低い山は実際は内容が異なる記事の対からなる。一方、類似度が高い山は同一内容の記事の対からなる。この分布は、2部グラフモデルを用いれば類似度に対して閾値を設けることにより同一内容の記事を精度良く対応づけることができることを表している。

4.3 閾値の最適化

図4に示した谷の周辺では同一内容の記事対とそうでない記事対が混じっている。このため、最も深い谷をそのまま閾値とすることはできない。最適な閾値を求めるため、類似度を0.3から0.6の間で変化させながら各日付毎に正解に対する対応付け結果のF値を最大にする閾値を求めた。F値は$F = \frac{2RP}{R+P}$により求めた。ここにRは再現率、Pは適合率である。最終的な評価を交差検定により行うため、記事を15日分に2等分した。以下、分割した記事群をA群およびB群と呼ぶ。等分した記事群の間では正解作成者が偏らないようにした。紙面の都合上、A群について求めたF値を図5に、最適な閾値とそのときの各日付毎に平均したF値を表2にそれぞれ示す。

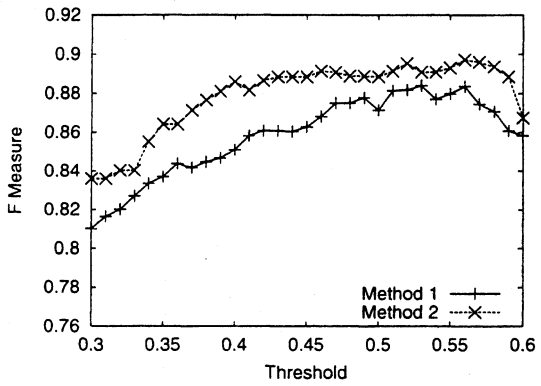


図5: F Mesure for Group A

表2: Optimum Thresholds

ベクトル		手法1	手法2
A群	閾値	0.53	0.56
	F値	0.88	0.90
B群	閾値	0.48	0.43
	F値	0.87	0.88

4.4 再現率、適合率、F値

表2に示したA(B)群の最適な閾値をB(A)群に適用し、各日付毎に平均した再現率、適合率、F値を表3に示す。

表3: Recall, Precision and F Measure

ベクトル		手法1	手法2
A群→B群	再現率	0.87	0.83
	適合率	0.87	0.88
	F値	0.86	0.85
B群→A群	再現率	0.92	0.94
	適合率	0.84	0.85
	F値	0.87	0.89

表2の結果より、最適化した閾値を未知の記事群に適用しても、0.85-0.89程度のF値が期待できる。1日当たりの正解記事対の数は平均10.7個のため、このF値は我々のシステムが正解記事対を約1-2個取りこぼす、ないしは正しくない記事対を約1-2個抽出することを示す。応用によっては、より高い性能が必要と考えられる。例えば問題として取り上げた新聞記事のフィルタリングなどの応用では、取りこぼしを防ぐ工夫が必要と考えられる。特徴ベクトルや類似度の改良によりさらに高い性能を得ることができるであろう。

5 まとめ

もともと2種類に分かれているような文書をクラスタリングではなく2部グラフマッチングにより効率良く対応づける手法について述べた。WWWニュース記事30日分を用いた実験より、F値による評価で0.85-0.89という高い性能を得た。今後は、2言語での対応文書検索や複数文書要約など、より広い分野への応用を考えている。

参考文献

[1] Peter F. Brown, Vincent J. Della Pietra, Peter V. deSouza, Jenifer C. Lai, and Robert L. Mercer. Class-based n-gram models of natural language. *Computational Linguistics*, Vol. 18, No. 4, pp. 283–298, 1992.

[2] Peter Cheeseman and David Wolpert. The AutoClass project, 1995. `http://ic.arc.nasa.gov/ic/projects/bayes-group/autoclass/`.

[3] N. Daishin. Hungarian method, 2000. `http://hp.vector.co.jp/authors/VA013417/hungaria.htm` (in Japanese).

[4] Inderjit S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. Technical Report 2001-05, UT CS, 2001.

[5] Noriko Kando. Overview of the second ntcir workshop. In *NTCIR Workshop 2*, pp. 35–44, Jul 2001.

[6] H. W. Kuhn. The Hungarian Method for the assignment problem. *Naval Research Logistics Quarterly*, pp. 83–97, 1955.

[7] Christopher D. Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*, chapter 14 Clustering. The MIT Press, 1999.

[8] 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸. 日本語形態素解析システム『茶筌』 version 2.0 使用説明書 第二版, 12 1999. NAIST Technical Report, NAIST-IS-TR99012. `http://chasen.aist-nara.ac.jp/`.