自 然 言 語 処 理 146－14
（２００１．１１．２１）

異なる発信元からのWWWニュース記事の内容に基づく対応付け

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

現在多くの新聞社等がWWWニュースサイトにて記事を配信している。配信される記事には同一内容のものがあるため、単純に記事を読むと同一内容の記事を複数回無駄に読んでしまう問題がある。これを避けるには、発信元が異なる同一内容の記事を対にする必要がある。具体的には、記事を発信元により分割し、2部グラフを構成する。その上で、2部グラフのマッチング問題を解くことにより対応付けを得る。本稿では、読売新聞と朝日新聞の記事を用いた対応付け結果を報告する。

Paring of WWW News Articles from Multiple Sources by Bipartite Matching

Seigo Tanimura†, Kumiko TANAKA-Ishii‡ and Hiroshi Nakagawa†

†Information Technology Center, The University of Tokyo

{tanimura,nakagawa}@r.dl.itc.u-tokyo.ac.jp

‡Interfaculty Initiative in Information Studies, The University of Tokyo

kumiko@ipl.t.u-tokyo.ac.jp

Many news paper companies submit their articles on the WWW. Because each company has its own site, one news story often appears in several sites. This forces news readers to read the identical news stories redundantly. In order to avoid this problem, we need to find articles with the same stories submitted from two distinct sources. Specifically, we model article relations with a bipartite graph. We then pair the same news stories by bipartite maching. We experimentally evaluate such matching on the articles submitted from Yomiuri Shimbun and Asahi Shimbun. The results show that our method pairs news articles at more than 86% of recall and 100% of precision at best.

1 はじめに

現在、新聞社など多くの発信元がWWWニュース記事を発信している。このため、読み手は多くの発信元を1箇所ずつ渡り歩きながらニュースを読まなければならない。その際、現在訪れている発信元にて読んだ記事と同一内容の記事が次の発信元に存在する可能性がある。この場合、そのような記事を読むのは無駄となることが多い。読み手が多くの発信元を渡り歩きながら無駄なく記事を読むためには、次に訪れる発信元にあるどの記事が現在訪れている発信元の記事と同一の内容であるかがわからなければならない。

読み手が現在訪れている発信元の記事と、次に訪れる発信元の記事のどれが同一の内容なのか、どのようにして求めればよいのだろうか? まず、現在訪れている発信元のある記事と同一内容の記事は、次に訪れる発信元に存在する。したがって、発信元によりノードを分割した2部グラフを用いて、記事群をモデル化することができると考えられる。それぞれの発信元に存在する同一内容の記事は、マッチングとして表現することができる。

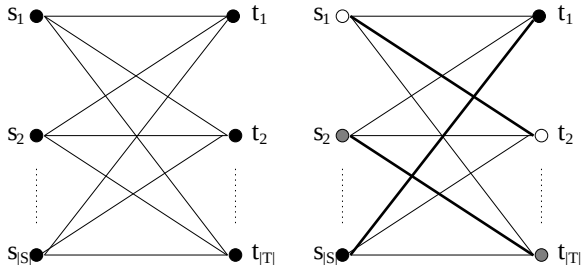
本稿では、異なる発信元から発信されたWWWニュース記事に含まれる同一内容の記事を、2部グラフのマッチングにより対応づける手法を提案する。以下、2節では、記事を2部グラフ、発信元が異なる同一内容の記事をマッチングによってそれぞれモデル化する。3節では、記事からの特徴ベクトル生成、マッチング、後処理の手法について述べる。4節では、実験に用いた記事データおよび実験結果について述べる。5節では、関連する研究や、今後の課

題について議論する。最後に 6 節にてまとめる。

2 モデル化

2.1 記事のモデル化

図 1(a) は、1 つのノードを 1 本の記事に見立てた 2 部グラフを図示したものである。ノードは発信元により左右に 2 分割されている。エッジは、左側の 1 ノードと右側の 1 ノードを結ぶもののみである。この 2 部グラフを用い、記事を以下のようにモデル化する。



(a) a Bipartite Graph (b) Matching in a Bipartite Graph

図 1: Modelling Articles by a Graph

2 部グラフを、$G(S,T,E)$ のように表す。ここに、S は左側、T は右側のノードの集合である。S および T を、式 (1) および (2) のようにそれぞれ定義する。

$$S = \{s_1, s_2, \cdots, s_i, \cdots, s_{|S|}\} \quad (1)$$

$$T = \{t_1, t_2, \cdots, t_i, \cdots, t_{|T|}\} \quad (2)$$

ここに、s_i や t_j は 1 つのノードである。

ここで、1 つのノードを 1 つの記事に対応させる。また、S および T を発信元に対応させる。これにより、s_i は S から発信された記事、t_j は T から発信された記事をそれぞれ表す。

E はノード間に張られたエッジである。1 本のエッジは、その両端にあるノードで表される記事が同一の内容である可能性があることを表す。E は一般には $S \times T$ の部分集合である。ここでは、S に含まれる任意の記事と T に含まれる任意の記事が同一の内容である可能性があるとする。これにより、E は式 (3) のようになる。

$$E = S \times T \quad (3)$$

以降では簡単のため、s_i と t_j を結ぶエッジ (s_i, t_j) を (i,j) と書く。

2.2 同一内容の記事対のモデル化

異なる発信元から発信された同一内容の記事対は、図 1(a) におけるマッチングとしてモデル化する。マッチングとはエッジの部分集合のうち、どのノードも高々1 回しか現れないものである。異なる発信元からの同一内容の記事を表すマッチングの例を、図 1(b) に示す。ここに、太線で描かれたエッジがマッチングである。色の濃さが同じノードは、同一内容の記事を表す。

3 手法

モデル化した記事に対し、マッチングを求めて同一内容の記事対を得る。しかし、このままではどのエッジをマッチングに含めれば同一内容の記事対をより多くマッチングに含められるのかわからない。このため、記事間の類似性を判断する尺度が必要となる。これを用いて、2 部グラフに重みを与える。その上で、重みつきマッチングを求める。最後に、マッチング結果に対して後処理を施す。

3.1 記事間の類似性尺度

記事間の類似性は、各記事について生成した特徴ベクトルの cosine 類似度によって測る。この方法は、文書の類似性を測るためによく用いられている。

以下、具体的な手法について述べる。まず、各記事を ChaSen 2.02 [18] により形態素解析し、単語に分割する。その上で、各記事の特徴ベクトルを生成する。同一内容の記事を求めるためには、特徴ベクトルは同一内容の記事の間では非常に高い類似度を、そうでない記事の間では非常に低い類似度を与えるものでなければならない。特徴ベクトルの生成手法としては、以下の 3 種類の手法を用いる。

手法 1. 名詞および未知語の頻度を特徴ベクトルの要素とする。

手法 2. 手法 1 にて、固有名詞および未知語の頻度を 2 倍にする。

手法 3. 名詞および未知語の連続を複合語とする。複合語、部分複合語、単語の頻度を特徴ベクトルの要素とする。

なお、部分複合語というのは、複合語に含まれる、

複合語よりも短い任意の長さの単語の連続である。例えば、複合語 ABC の部分複合語は、AB、BC、A、B、C である。これにより、複合語の頻度が単語の頻度よりも少なく数えられることを防ぐ。

3.2 記事間の類似度

記事間の類似度は、各記事の特徴ベクトルの cosine 類似度により求める。具体的には、式 (4) の $\cos\theta$ を類似度とする。ここに、$\mathbf{x}_1$ および $\mathbf{x}_2$ は各記事の特徴ベクトルである。

$$\cos\theta = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{|\mathbf{x}_1| \cdot |\mathbf{x}_2|} \tag{4}$$

3.3 2 部グラフの重みつきマッチング

2 部グラフに重みを与える方法としては、以下の 2 通りがある。重みの与え方により、解法が異なる。

- 各 s_i および t_j について、T および S のノードをそれぞれマッチングに含めたい順に順位付ける。

  この場合、重みつきマッチングは安定結婚問題を解くことにより求まる。ただし、安定結婚問題は $|S| = |T|$ が成り立たなければ定義できない。一般に発信される記事の数は発信元によって異なる。このため、我々の手法では安定結婚問題は用いない。

- 各 (i,j) について、エッジの重み $w(i,j)$ を与える。

  この場合、重みつきマッチングは Hungarian method [2, 8] により解ける。Hungarian method とは、$w(i,j) > 0$ となる 2 部グラフにおいて、エッジの重みの総和 W_{match} が最大となるようなマッチング (最大重みマッチング) を求める。W_{match} は式 (5) のように表せる。

$$W_{match} = \sum_{(i,j)\in M} w(i,j) \tag{5}$$

  ここに、$M \subseteq E$ はマッチングである。安定結婚問題と異なり、$|S| \neq |T|$ であっても解くことができる。このため、我々は Hungarian method を用いてマッチングを求める。

  Hungarian method の時間計算量は、$O\left((|S|+|T|)^3\right)$ となることが知られている。一見すると、この計算量はかなり多そうに見える。しかし、後述する実験に用いた記事数の範囲では、計算時間は大きな問題とはならなかった。

3.4 後処理

2 部グラフのマッチングだけでは記事群を取り扱うに十分ではない。現実には、ある発信元が同一内容の記事を 2 回配信し、片やもう片方の発信元は 1 回しか記事を配信しない可能性もある。この場合、グラフのマッチングだけではこれらの 3 記事をすべて同一記事として対応づけることはできない。このため、各発信元毎に同一内容の記事を併合する必要がある。

一見すると、これはクラスタリングの問題に見える。一般に、クラスタリングを行うことにより得られるクラスタはクラスタ数やクラスタの種として用いる記事の選択などに依存する。このため、最適なクラスタリングを行うことは難しい。しかし、ここではクラスタの種としてマッチングにより対応付けがとれた記事を用いる。したがって、一般的なクラスタリングよりも問題を簡単化できる。

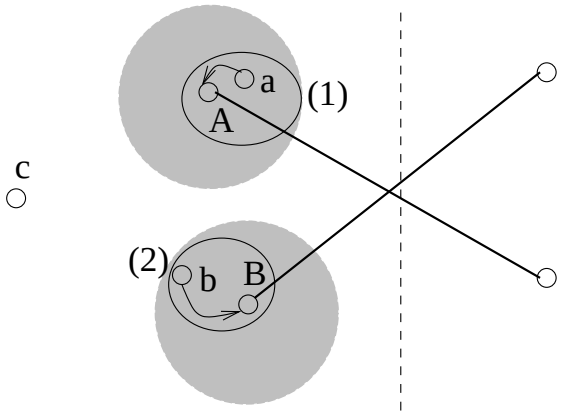


図 2: Postprocessing of Articles from a Single Source

記事を併合する手法を図 2 に示す。太線は 3.3 節の手法により得たマッチングである。A と B はマッチングに含まれる記事である。a、b、c はマッチングに含まれない記事である。記事間の距離の近さが類似度の高さを表す。灰色の領域は、併合の対象とする記事の範囲である。併合は、マッチングに含まれる記事と灰色の領域に含まれかつマッチングに含まれない記事の組について、類似度が高いものから順にまとめることにより行う。灰色の領域に記事ががなくなった時点で、併合を終了する。図 2 では、まず A と a の類似度が最も高いため、これらを併合する (1)。続いて、類似度が 2 番目に高い B と b を同

様に併合する (2)。c は灰色の領域に入らないため、そのまま残される。以上の後処理を、2 つの発信元それぞれの記事について行う。

4 実験

4.1 データ

実験には読売新聞および朝日新聞が WWW にて発信している社会面記事を用いた。日付は 8 月 24 日付から 28 日付までの 5 日分とした。記事数を表 1 に示す。

表 1: Number of News Articles for 5 days

発信元	24 日	25 日	26 日	27 日	28 日
読売	56	46	35	53	60
朝日	41	25	23	32	34

4.2 類似度の分布

予備実験として、各日付毎に読売の記事と朝日の記事間の類似度を求めた。その上で、3.1 節で述べた各特徴ベクトル生成手法について類似度の分布を調べた。求めた類似度の分布を図 3 に示す。

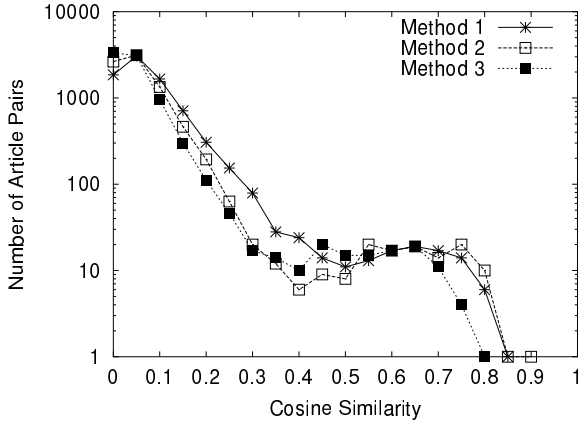


図 3: Distribution of Cosine Similarity

類似度は 0.05 付近に集中した。その結果、99%以上の記事対が類似度 0.5 以下の範囲内に含まれた。これらは異なる内容の記事対である。また、類似度が 0.5 から 0.8 の間に 100 個程度の記事対が集中した。これらは同一内容の記事対である。

4.3 同一記事の対応付け

各日付毎に読売と朝日の記事について最大重みマッチングを求めた。その上で、マッチングにより得た記事対の類似度について、分布を調べてみた。類似度の分布を図 4 に示す。

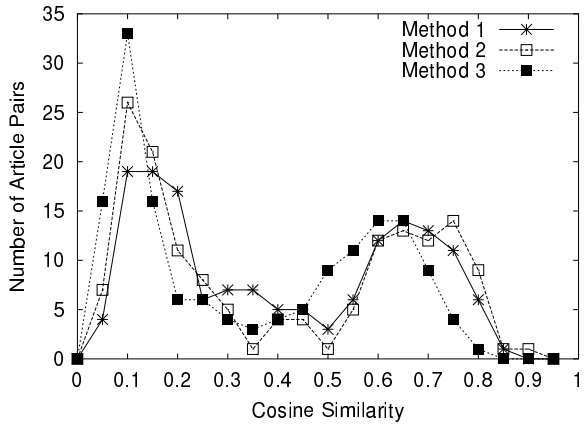


図 4: Distribution of Cosine Similarity for Matched Article Pairs

異なる内容の記事対数が減り、同一内容の記事対が目立つようになった。ベクトル生成手法ごとに結果を見ると、手法 2 が同一内容の記事対に対して手法 1 よりも若干高い類似度を与えた。一方、手法 3 は手法 1 に比べ、同一内容の記事対に対して 0.1 程度低い類似度を与えた。これは、同じ意味の複合語でも発信元により表記が変わることがあるためである。特徴ベクトルに含まれていた複合語および部分複合語の例を図 5 に示す。なお、読売での表記は「平田隆志さん (51)」、朝日では「平田隆志警部補 (51)」である。

(a) Yomiuri
平田
平田 隆志
平田 隆志 さん
平田 隆志 さん (51)
隆志
隆志 さん
隆志 さん (51)
さん
さん (51)
(51)

(b) Asahi
平田
平田 隆志
平田 隆志 警部補
平田 隆志 警部補 (51)
隆志
隆志 警部補
隆志 警部補 (51)
警部補
警部補 (51)
(51)

図 5: Example of Compound Words in a Vector

表 2: Incorrect Article Pairs

日付	24 日		25 日		26 日		27 日		28 日	
後処理	なし	あり	なし	あり	なし	あり	なし	あり	なし	あり
手法 1	0	0	0	0	1	**0**	0	0	0	0
手法 2	0	0	0	0	0	0	0	0	0	0
手法 3	1	**0**	1	**0**	2	2	0	0	0	0

表 3: Recall and Precision with Postprocess

日付	24 日		25 日		26 日		27 日		28 日	
ベクトル生成手法	rec.	prec.	rec.	prec.	rec.	prec.	rec.	prec.	rec.	prec.
1	1.00	1.00	0.92	1.00	1.00	1.00	0.89	1.00	0.81	1.00
2	1.00	1.00	0.92	1.00	1.00	1.00	0.94	1.00	0.86	1.00
3	1.00	1.00	0.83	1.00	0.83	0.83	0.83	1.00	0.71	1.00

単語が 1 語異なるだけでも、異なる部分複合語が多く生成されてしまう。このため、手法 3 では類似度が落ちたと考えられる。

続いて、マッチングにより得られた記事対から閾値以上の類似度を持つものを同一内容の記事対として抽出した。閾値は 0.5 とした。次に、人手により正解記事対をマッチングとして作成した。その上で、同じ発信元が同一内容の記事を複数回発信したため、抽出された記事対に含まれた不正解記事対の数を調べた。また、抽出された記事対に対して後処理を施した場合の不正解記事対数も求めた。不正解記事対の数を表 2 に示す。後処理を施した場合については、マッチングにより得た記事対に含まれる各記事が正解に含まれる各記事を併合していれば正解とした。太字は改善した結果である。

26 日付の記事に対して手法 3 を用いた場合を除き、後処理の効果が現れている。26 日付の記事に対する手法 3 の結果では、読売と朝日が同一の記事をそれぞれ 2 回配信していた。これらの記事はマッチングに含まれてしまったため、後処理による改善ができなかったと考えられる。

さらに、後処理により得た同一記事対について recall および precision を用いて評価した。recall および precision はそれぞれ式 (6) および (7) のように定義した。ここに、E は抽出された記事対の集合、C は正解記事対の集合である。ただし、$|E \cap C|$ には併合された記事が正解に含まれていた場合も含めている。

$$recall = \frac{|E \cap C|}{|C|} \tag{6}$$

$$precision = \frac{|E \cap C|}{|E|} \tag{7}$$

precision の低下は後処理の効果がなかったためである。recall の低下は、マッチング結果から記事対を抽出するために用いた類似度の閾値 0.5 が一部の日付については最適でなかったためと考えられる。対策として、記事群に対する閾値の動的な決定を検討している。

5 議論

5.1 関連研究

文書群をグラフにより取り扱う研究例としては、Uramoto らの研究 [15] がある。彼らはプッシュ型情報配信システム向けに、有向グラフを用いて関連性のある文書群を時系列順に記述している。文書間の時系列はエッジの向きとして表現している。我々の手法では、時系列は考慮しない。また、グラフは無向である。このため、グラフのトポロジはかなり異なったものとなる。

Watanabe らは、彼らは TV ニュース記事 143 件と新聞記事 100 件の対応付けをとる研究を行っている [16]。TV ニュースからはテロップを認識することによりテキストを得ている。TV および新聞を複数の異なる発信元と見なせば、我々の研究によく似ている。ただし、彼らの手法では、ある程度類似している記事の組合せはすべて対応する可能性があるとしている。我々の研究は、記事の組合せをマッチングにより最適化している点で Watanabe らの研究とは異なる。

また、文書の集合が与えられた時に、同一内容の文書をまとめる文書クラスタリングの技術がある。文書クラスタリングに関する多くの研究例では、新聞記事 [4–7, 12, 13] ないしは WWW 上の文書 [1, 3, 9,

11, 17] を対象としている。新聞記事を対象とした研究では、特に大量の文書を必要とするのでない限り、発信元を 1 箇所に限定することが多い。これは、発信元が異なることにより記事の表記に差異が現れることを防ぐためである。同様の理由から、記事の分野 (政治、経済、社会など) についても限定することが多い。異なる分野が混在する記事を対象とする場合もあるが、その場合には記事から分野を自動的に得ることが目的になることが多い [14]。このように、新聞記事を対象としたクラスタリングでは、対象とする記事について発信元や分野などの性質を一貫させるか、対象記事における性質の違いを自動的に獲得するすることのいずれかを目的とする。WWW 上の文書については、発信元や分野がどのようなものかがよくわかっていない。このため、WWW 上の文書をクラスタリングの対象とする場合、発信元や分野は限定しないことが多い。我々の知る限り、異なる性質、すなわち発信元の記事を対象とし、かつ対象記事における性質の違いを獲得することを目的としない研究はまだ例がない。

5.2 今後の課題

1. ニュースブラウザの作成
   応用として、異なる複数の発信元を渡り歩きながら、同一内容の記事を自動的にまとめて表示するようなニュースブラウザの作成を考えている。
2. 他国語 WWW ニュース記事への拡張
   先日のアメリカでのテロ事件の際、アメリカでは報道されたが日本では報道されなかったニュースがいくつかあった。我々の手法はそのような国内で見過ごされているニュース記事をいち早く発見するために有用と考えられる。他国語への拡張の際には単語などの翻訳が必須となる。辞書を用いれば翻訳は可能なように見える。しかし、ニュース記事では辞書にない新しい単語を用いる可能性がある。これを解決するため、辞書の自動拡張 [10] や、WWW 上からの up-to-date な情報の利用などを併用する必要があると考えられる。

6 まとめ

WWW ニュース記事に含まれる同一内容の記事を 2 部グラフマッチングにより対応づける手法について述べた。5 日分の記事を用いた実験より、名詞および未知語の頻度を要素とし、固有名詞および未知語の頻度を 2 倍にした特徴ベクトルではほぼ完全な対応付けを得た。現在、1 ヶ月分の記事を対象とした、より大規模な評価を行なっているところである。

参考文献

[1] P. G. Anick and S. Vaithyanathan. Exploiting clustering and phrases for Context-Based information retrieval. In *ACM SIGIR97*, pp. 314–323, Jul 1997.

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

[3] K. Eguchi. Adaptive cluster-based browsing using incrementally expanded queries and its effects. In *ACM SIGIR99*, pp. 265–266, Aug 1999.

[4] F. Fukumoto and Y. Suzuki. An automatic clustering of articles using dictionary definitions. In *COLING96*, Vol. 1, pp. 406–411, Aug 1996.

[5] V. Hatzivassiloglou, L. Gravano, and A. Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In *ACM SIGIR2000*, pp. 224–231, Jul 2000.

[6] M. A. Hearst and J. O. Pedersen. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *ACM SIGIR96*, pp. 76–84, Aug 1996.

[7] T. Ikeda, A. Okumura, and K. Muraki. Information classification and navigation based on 5W1H of the target information. In *COLING-ACL98*, Vol. 1, pp. 571–577, Aug 1998.

[8] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, pp. 83–97, 1955.

[9] M. Mechkour, D. J. Harper, and G. Muresan. The WebCluster project. Using clustering for mediating access to the world wide web. In *ACM SIGIR98*, pp. 357–358, Aug 1998.

[10] G. Petasis, A. Cucchiarelli, P. Velardi, G. Paliouras, V. Karkaletsis, and C. D. Spyropoulos. Automatic adaptation of proper noun dictionaries through cooperation of machine learning an probabilistic methods. In *ACM SIGIR2000*, pp. 128–135, July 2000.

[11] D. Roussinov, K. Tolle, M. Ramsey, and H. Chen. Interactive internet search through automatic clustering: an empirical study. In *ACM SIGIR99*, pp. 289–290, Aug 1999.

[12] H. Schütze and C. Silverstein. Projections for efficient document clustering. In *ACM SIGIR97*, pp. 74–81, Jul 1997.

[13] C. Silverstein and J. O. Pederson. Almost-Constant-Time clustering of arbitrary corpus subsets. In *ACM SIGIR97*, pp. 60–66, Jul 1997.

[14] N. Slonim and N. Tishby. Document clustering using word clusters via the information bottleneck method. In *ACM SIGIR2000*, pp. 208–215, Jul 2000.

[15] N. Uramoto and K. Takeda. A method for relating multiple newspaper articles by using graphs, and its application to webcasting. In *COLING-ACL98*, Vol. 2, pp. 1307–1313, Aug 1998.

[16] Y. Watanabe, Y. Okada, K. Kaneji, and M. Nagao. Aligning articles in TV newscasts and newspapers. In *COLING-ACL98*, Vol. 2, pp. 1381–1387, Aug 1998.

[17] O. Zamir and O. Etzioni. Web document clustering: A feasibility demonstration. In *ACM SIGIR98*, pp. 46–54, Aug 1998.

[18] 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://cl.aist-nara.ac.jp/lab/nlt/chasen/`.