2E1-01

語の共起情報を用いた文書クラスタリング

Document clustering that uses co-occurrence information on word

小熊淳一　　内海彰
Junichi Oguma　Akira Utsumi

電気通信大学電気通信学研究科
Department of Systems Engineering, The University of Electro-Communications.

In this paper, we propose two methods (co-occurrence-based method and term-importance-based method) of generating document vectors used for document clustering techniques. A co-occurrence-based method uses both term frequency and term co-occurrence frequency. A term-importance-based method calculates the degree of importance using a co-occurrence graph of a document, and then attaches such importance value to the tfidf value of a term in the document vector. This paper also reports an evaluation experiment in which the performance of the proposed methods is compared with that of the existing tfidf-based method.

1. はじめに

今日，Web 等を利用して電子化された文書を容易に入手できるようになった．そのため，ユーザーが必要とする文書を効率的に得るための技術として文書クラスタリングが注目されており，ユーザーの意図に応じて分類結果を変えることができる検索支援システム [1] や，適合フィードバックにより得られるクエリの情報をクラスタリングにおける観点として活用する手法 [2] などの研究が行われている．

文書クラスタリングは一般的に，繰り返し言及される語句はその文書における重要な概念を表すという仮定のもとに，名詞の出現頻度の情報を用いて行われる．しかし，人が文書を分類する場合には，名詞単独で見るだけでなくその名詞と同一文中に出現する他の名詞との結び付きの強さが重要となることがある．また，出現頻度が低い名詞であっても文書内で重要な概念となるものが存在し，人はその概念に基づいて文書の分類を行うことがある．このような場合には名詞の出現頻度の情報のみでは人手による分類と同等以上の分類をシステムで行うことができない．

そこで本研究では，文書ベクトルの構成法として新たに，語の共起回数と出現頻度を考慮する方法と，語の共起に基づいて抽出したキーワード [3] を用いた手法を提案し，その有効性を検討する．

2. ベクトル空間モデルに基づく文書クラスタリング

2.1 文書ベクトルの構成

ベクトル空間モデルに基づく文書クラスタリングでは，各文書を文書ベクトルと呼ばれる指標で表現する．

既存の研究の多くでは，文書 D_i における文書ベクトル d_i の各属性値 $d_i^k (k = 1, .., n)$ には，次式に示す D_i における tf・idf 値が用いられる．

$$d_i^k = tf_i(\lambda^k) \cdot idf(\lambda^k) \quad (1)$$

$$tf_i(\lambda^k) = \frac{D_i\text{中の名詞}\lambda^k\text{の頻度}}{D_i\text{中の全名詞の出現頻度の総和}} \quad (2)$$

$$idf(\lambda^k) = \log\left(\frac{\text{総文書数}}{\text{名詞}\lambda^k\text{が出現する文書数}}\right) + 1 \quad (3)$$

なお各ベクトル d_i は長さが 1 となるように正規化される．

2.2 階層的クラスタリング

文書 D_i，D_j の文書ベクトルを d_i，d_j とすると，一般的にそれらの類似度 $s(d_i, d_j)$ は d_i と d_j のなす角の余弦で定義される．

$$s(d_i, d_j) = \frac{d_i \cdot d_j}{|d_i|\,|d_j|} \quad (4)$$

階層的クラスタリングでは，(4) 式により得られる文書間類似度を用いて以下のアルゴリズムによりクラスタを構成する．

1. 初期設定として，個々の文書をそれぞれ 1 つのクラスタとする．
2. 類似度が最大のクラスタ対を結合する．クラスタ数が 1 になれば，終了する．
3. 全てのクラスタについてクラスタ間の類似度を再計算して 2 に戻る．

クラスタ G_i，G_j 間の類似度には，G_i に属する任意の文書の文書ベクトル x と G_j に属する任意の文書の文書ベクトル y の類似度のうち最小のものをそのクラスタ間の類似度とする最長距離法を用いる．

$$s(G_i, G_j) = \min_{x \in G_i, y \in G_j} s(x, y) \quad (5)$$

2.3 非階層的クラスタリング

非階層的クラスタリング手法として，k-means 法を用いる．k-means 法では，以下のアルゴリズムによりクラスタを構成する．クラスタ分割数は K とし，文書数は m，クラスタ W に含まれる文書数は $N(W)$ とする．

1. K 個の初期クラスタを適当に決める．
2. (6) 式により得られる文書 D_i をクラスタ W に移動したときの誤差の増加量 $e(i, W)$ を計算し，その最小値を持つクラスタ W に文書 D_i を移す．

$$e(i, W) = \frac{N(W)D(i, W)^2}{N(W) + 1} - \frac{N\{W(i)\}D\{i, W(i)\}^2}{N\{W(i)\} - 1} \quad (6)$$

連絡先: 小熊淳一，電気通信大学電気通信学研究科，東京都調布市調布ヶ丘 1-5-1 電気通信大学　電気通信学部システム工学科　内海研究室，0424-43-5258，og@utm.se.uec.ac.jp

ただし，$D(i,W)$ は文書 D_i とクラスタ W との距離であり，次式により定義される．

$$f_W^k = \frac{\sum_{x \in L_i} d_x^k}{N(W)} \quad (7)$$

$$D(i,W) = \sum_{k=1}^{n} \left\{ d_i^k - f_W^k \right\} \quad (8)$$

3. あるクラスタから他のクラスタへの文書の移動がなければ終了する．そうでなければ 2. に戻る．

3. 提案する文書ベクトルの構成法

3.1 共起回数と単語頻度とを併せた文書ベクトル

文書集合中に出現する名詞の集合 $\Lambda = \left\{ \lambda^1, .., \lambda^n \right\}$ に対して，任意の 2 つの名詞 λ^k，λ^l の文書 D_i における出現頻度を f_i^k，f_i^l として，共起 tf・idf 値 $d_{co,i}^{k,l}$ を次式により定義する．

$$d_{co,i}^{k,l} = tf_{co,i}(\lambda^k, \lambda^l) \cdot idf_{co}(\lambda^k, \lambda^l) \quad (9)$$

$$tf_{co,i}(\lambda^k, \lambda^l) = \begin{cases} \dfrac{D_i \text{ 中の名詞 } \lambda^k,\ \lambda^l \text{ が共起する文の数}}{D_i \text{ の文の総数}} \\ \qquad (f_i^k \geq f_0 \text{かつ } f_i^l \geq f_0) \\ 0 \quad (f_i^k < f_0 \text{または } f_i^l < f_0) \end{cases} \quad (10)$$

$$idf_{co}(\lambda^k, \lambda^l) = \log \frac{\text{総文書数}}{S(\lambda^k, \lambda^l)} + 1 \quad (11)$$

ただし，$S(\lambda^k,\lambda^l)$ は文書集合中で $tf_{co,i}(\lambda^k, \lambda^l) > 0$ となる文書数である．なお各ベクトル $d_{co,i}$ は長さが 1 になるように正規化される．そして，(1) 式の tf・idf 値 d_i^k とあわせて，(12) 式に示す文書ベクトル $d_{ct,i}$ を構成する．

$$d_{ct,i} = ((1-\alpha)d_i^1, .., (1-\alpha)d_i^n, \alpha d_{co,i}^{1,2}, .., \alpha d_{co,i}^{n-1,n}) \quad (12)$$

なお α は 0 から 1 までの値を取る定数であり，共起 tf・idf 値の影響の度合を表す．

3.2 キーワードに基づく文書ベクトル

文書から得られる語どうしの結び付きの関係から複数の語どうしを直接または間接的につなぐ働きをする語はその文書において重要性が高いという仮定に基づき，文書の共起情報からグラフを構成し語の重要度を計算する [3]．次にその重要度を用いて (1) 式の tf・idf 値 d_i^k に重みづけをして文書ベクトルを構成する．

3.21 共起グラフ

文書 D_i の共起グラフ C_i を以下の手順で生成する．まず，文書 D_i 中に f_0 回以上出現する名詞をノードとして取り出す．次に任意の二つの名詞 λ^j，λ^k について次式に示す *Jaccard* 係数値 を計算する．そしてその値の高いものから 1 ノードあたりのリンク数の平均 a が規定値 a_0 となるまでリンクを張る．

$$Jaccard \text{ 係数値} = \frac{\text{名詞}\lambda^j,\lambda^k\text{両方を含む文の数}}{\text{名詞}\lambda^j,\lambda^k\text{の少なくとも一方を含む文の数}} \quad (13)$$

$$a = \frac{\text{総リンク数}}{\text{総ノード数}} \quad (14)$$

3.22 語の重要度

共起グラフ C_i における語 λ^k の重要度を以下の方法で計算する．ただし，C_i は m 個の連結していないサブグラフ $c_i^1, .., c_i^m$ の集合である．また c_i^j に含まれるノードの集合を N_i^j とする．

ノード q,r に対して拡張パス長 $p\prime(q,r)$ を次のように定義する．

$$p\prime(q,r) = \begin{cases} p(q,r) & \text{ノード } q,r \text{ 間にリンクが存在する場合} \\ \sum_{s=1}^{m} \max_{x,y \in N_i^j} p(x,y) & \text{それ以外の場合} \end{cases} \quad (15)$$

ただし，$p(q,r)$ は共起グラフ C_i におけるノード q とノード r の最短パスの長さである．

次に名詞 λ^k に対応するノード k について，ノード k を取り除いたグラフと元のグラフとのパス長の平均 L の差を考え，これを名詞 λ^k の文書 D_i における重要度 CB_i^k とする．

$$CB_i^k = L_{G_k} - L_k \quad (16)$$

ただし，L_k は，ノード k を除く全てのノードの組についての $p\prime(q,r)$ の平均値，L_{G_k} は，ノード k を取り除いたグラフにおける $p\prime(q,r)$ の平均を表す．

3.23 重要度に基づく文書ベクトルの生成

文書 D_i の文書ベクトル $d_i\prime = (d_i^1\prime, .., d_i^n\prime)$ の各要素 $d_i^k\prime$ を次式で定義する．

$$d_i^k\prime = d_i^k \cdot \left(1 + \beta \sum_{i=1}^{N} CB_i^k \right) \quad (17)$$

ただし，d_i^k は (1) 式で定義された tf・idf 値，CB_i^k は文書 D_i における名詞 λ^k の重要度，N は総文書数を表す．β は 0 以上の値をとる定数であり，文書ベクトルへの重要度の影響の度合を表す．

4. 評価

初めに検索サイトでクエリを入力し，そのクエリに関連する文書として得られた HTML 文書，論文，ニュース記事の本文のみを各 50 件ずつ用意し，人手で分類を行い正解のクラスタの集合を作成した．次に (1) 式による tf・idf 値を用いた文書ベクトル（単語頻度法），共起回数と単語頻度とを併せた文書ベクトル（共起頻度法）および名詞の重要度を用いた文書ベクトル（重要語強調法）をそれぞれ用いて階層的クラスタリングを行い，正解と比較した．

本研究で得られるクラスタと正解クラスタがどの程度近いかの指標として，以下の評価基準を用いる．クラスタ A，B に属する文書のうち A，B に共通する文書の数を $O(A,B)$ と定義する．まず，正解のクラスタの集合 $g = \{G_1, .., G_c\}$，各手法によるクラスタの集合 $g\prime = \{G\prime_1, .., G\prime c\}$ を得る．$O(G_t,G\prime_u)$ $(\forall G_t \in g, \forall G\prime_u \in g\prime)$ が最大となる G_t を正解クラスタ 1，$G\prime_u$ を best クラスタと呼ぶ．次に，g から正解クラスタ 1 を取り除いたクラスタの集合 $h = \{H_1, .., H_{c-1}\}$，$g\prime$ から best クラスタを取り除いたクラスタの集合 $h\prime = \{H\prime_1, .., H\prime_{c-1}\}$ を得る．$O(H_v,H\prime_w)$ $(\forall H_v \in h, \forall H\prime_w \in h\prime)$ が最大となる H_v を正解クラスタ 2，$H\prime_w$ を best2 クラスタと呼ぶ．そして best クラスタ及び best2 クラスタの再現率，適合率，及び F 値を用いて本研究の手法が既存の手法に比べ，人手での分類に近い分類をするのに有効であるかを評価する．

$$\text{再現率 } (R) = \frac{best(best2) \text{ クラスタ中の適合文書数}}{\text{正解クラスタ 1（正解クラスタ 2）の文書数}} \quad (18)$$

表 1: 最長距離法による HTML 文書における各手法の性能

	best クラスタ			best2 クラスタ		
	単語頻度法	共起頻度法	重要語強調法	単語頻度法	共起頻度法	重要語強調法
再現率	0.800	0.667	0.867	0.182	1.000	0.727
適合率	0.667	0.645	0.812	0.286	1.000	0.727
F 値	0.727	0.656	0.839	0.222	1.000	0.727

表 2: 最長距離法による論文における各手法の性能

	best クラスタ			best2 クラスタ		
	単語頻度法	共起頻度法	重要語強調法	単語頻度法	共起頻度法	重要語強調法
再現率	1.000	0.900	0.800	0.700	0.667	0.455
適合率	0.733	0.450	0.400	0.636	0.857	0.500
F 値	0.846	0.600	0.533	0.667	0.750	0.476

表 3: 最長距離法によるニュース記事における各手法の性能

	best クラスタ			best2 クラスタ		
	単語頻度法	共起頻度法	重要語強調法	単語頻度法	共起頻度法	重要語強調法
再現率	0.400	0.400	0.650	0.350	0.750	0.750
適合率	0.615	0.615	0.684	0.778	1.000	1.000
F 値	0.485	0.485	0.667	0.483	0.857	0.857

表 4: k-means 法による HTML 文書における各手法の性能

	best クラスタ			best2 クラスタ		
	単語頻度法	共起頻度法	重要語強調法	単語頻度法	共起頻度法	重要語強調法
再現率	0.467	0.300	0.500	0.545	0.636	0.545
適合率	0.609	1.000	0.882	0.857	0.778	1.000
F 値	0.528	0.462	0.638	0.667	0.700	0.706

表 5: k-means 法による論文における各手法の性能

	best クラスタ			best2 クラスタ		
	単語頻度法	共起頻度法	重要語強調法	単語頻度法	共起頻度法	重要語強調法
再現率	0.800	1.000	1.000	1.000	0.636	0.636
適合率	1.000	1.000	0.769	0.250	0.292	0.304
F 値	0.889	1.000	0.870	0.400	0.400	0.412

表 6: k-means 法によるニュース記事における各手法の性能

	best クラスタ			best2 クラスタ		
	単語頻度法	共起頻度法	重要語強調法	単語頻度法	共起頻度法	重要語強調法
再現率	0.350	0.600	0.350	0.444	0.444	0.667
適合率	0.636	0.667	1.000	0.500	1.000	0.333
F 値	0.452	0.632	0.519	0.471	0.615	0.444

$$\text{適合率 } (P) = \frac{best(best2) \text{ クラスタ中の適合文書数}}{best(best2) \text{ クラスタ中の文書数}} \quad (19)$$

$$F \text{ 値} = \frac{2RP}{R+P} \quad (20)$$

それぞれの文書集合において，各文書ベクトルを用いて最長距離法及び k-means 法を行い，best クラスタ，best2 クラスタの再現率，適合率，F 値を表 1〜6 に示す．

5. 考察

5.1 共起頻度法の有効性

表 1 から表 3 の結果から，文書間類似度を計ってクラスタリングした結果，単語頻度法に比べて共起頻度法は全体的に F 値が低かった．その理由として，正解クラスタ 1（正解クラスタ 2）に属する文書のみに共通する名詞間の共起が少なく，文書集合全体に出現する名詞間の共起が文書間類似度に大きく影響してしまったことが挙げられる．また，best2 クラスタの F 値は高くなっているが，これは best クラスタに多くの文書が集まってしまい，best2 クラスタに含まれる文書が少ないことが挙げられる．

表 4 から表 6 の結果から，クラスタ中心からの誤差を考える方法では，論文とニュース記事において単語頻度法に比べて共起頻度法は F 値が高くなった．その理由として，文書集合全体に出現する名詞間の共起の影響よりも同じ正解のクラスタ内に属する文書に共通して出現する共起の影響が大きくなったことが挙げられる．

以上から，共起頻度法は，文書間類似度をその文書ベクトルのなす角の余弦により定義する方法では人手に近い分類をするのに有効ではないが，k-means 法のようにクラスタ中心からの誤差を考える方法では人手に近い分類を与えるのに有効であるという結果となった．

5.2 重要語強調法の有効性

表 1 から表 6 の結果から，単語頻度法に比べて重要語強調法は F 値が高くなる傾向にあり，文書集合の種類によっては人手に近い分類をするのに有効であることを示している．

論文における結果をみると，表 2 から best クラスタ，best2 クラスタともに F 値が低くなった．その理由として重要度の高い名詞が複数の文書に登場することが少なかったことと，文書集合全体で出現しやすい名詞の重要度が高くなってしまったことが挙げられる．

また，HTML 文書やニュース記事の場合のように文書集合全体に共通して出現頻度の多い名詞が存在し，正解クラスタ 1（正解クラスタ 2）に属する多くの文書に共通するキーワードが存在する場合には，本手法は単語頻度法に比べ，人手による文類に近い分類をするのに有効であるといえる．

6. おわりに

本研究では，文書ベクトルの構成法として，語の共起回数と出現頻度を考慮する方法と，語の共起に基づいて抽出したキーワードを用いた手法を提案し，人手での分類に近い分類を与えるのに有効であることを示した．

今後の課題としては，名詞そのものの共起だけでなく類義語との共起関係から類似度を計算したり，キーワードが出現する文書数から使用するキーワードの選別をすることにより，分類精度を挙げることが考えられる．

参考文献

[1] 折原 大：文書クラスタリングを用いた Web 検索支援システム，電気通信大学電子情報学専攻修士論文 (2002).

[2] 江口 浩二，伊藤 秀隆，隈元 昭，金田 彌吉：漸次的に拡張されたクエリを用いた適応的文書クラスタリング法，電子情報通信学会論文誌，Vol.J82-D-I，No.1，pp.140-149 (1999).

[3] 松尾 豊，大澤 幸生，石塚 満：Small World 構造に基づく文書からのキーワード抽出，情報処理学会論文誌，Vol.43, No.6, pp.1825–1832 (2002).