E-059

単語間の関係に基づくWeb文書クラスタリング
Web document clustering based on the relations between words

仁科 朋也[1]　内海 彰[2]
Tomoya Nishina　Akira Utsumi

1 はじめに

今日，Webからユーザーの望む情報を得る手段としてGoogleなどのサーチエンジンが一般的に利用される．しかし，ユーザーが望む情報を持たないページも多数表示されるため，各ページがユーザーの望む情報を含むかどうかを判断するのに時間と労力を割かなければならない．このような負担を軽減するための検索支援手法として，検索結果を分類して表示するWeb文書クラスタリングが挙げられる．クラスタリング方法にはWebページに含まれるテキスト（文章）を用いる方法 [1-3] や，HTMLの構造を用いる方法 [4]，Webページのリンクを用いる方法 [5] など様々な方法がある．

その中でも文章を用いる方法が一般的であり，検索結果のタイトルとスニペットを扱う手法 [1,2] と文書全体を扱う手法 [3] に分かれる．前者は後者に比べ，取得時間や処理時間が短く，ノイズとなる単語が混ざりにくいなどの利点がある．これらの手法 [1,2] では，単語の出現頻度や出現位置から重要語を選び，その語を含む文書をクラスタとする非排他的クラスタリング（文書を複数のクラスタに割り振ることを許すクラスタリング）を行っている．しかし，この重要語の選び方では単語間の意味的な関係を考慮していないので，重要語どうしが類似した意味を表していると，類似したクラスタを出力してしまう．

そこで単語間の意味の関係を考慮した重要語を選ぶことによって，類似したクラスタを極力出力せず，文書を分類できると考えられる．本研究ではこの考えに基づいた非排他的クラスタリング手法を提案し，論文 [1,2] の手法との比較評価を通じて本手法の有効性を示す．

2 提案手法の概要

本研究の手法の概要を以下に示す．

1. クエリを受け取り，Googleによる検索結果のタイトルとスニペットを文書として取得する．
2. 文書に対し形態素解析を行う．なお形態素解析には茶筌[3]を用いる．
3. 形態素解析で名詞・英字と判断された単語を，整形して抽出する．
4. 抽出した単語から重要語をn語取り出す．
5. 重要語をクラスタ名とするクラスタを作成し，文書中にその重要語が含まれていればそのクラスタにその文書を含める．なおどのクラスタにも含まれない文書を「その他」のクラスタに分類する．

3 形態素解析結果からの名詞抽出

概要の手順3の整形方法の詳細を述べる．まず形態素解析により名詞及び英字と判断された単語を抽出する．なお非自立の名詞や代名詞は除き，英字が連続している場合はつなげて1つの名詞とする．また各単語 W_i に対し $df(W_i)$ 値（単語 W_i を含む文書の数）を計算し，一定値以下の単語を除外する．これは $df(W_i)$ 値が極端に低い単語を重要語とすると，クラスタに含まれる文書数も極端に低くなるためである．さらにクエリ及びクエリの一部となる単語は，ほぼ全ての文書に出現するため，手順4の重要語抽出に大きな影響を及ぼすので取り除く．

次に形態素解析による単語の誤りを修正する．すべての単語の組合せに対して以下のアルゴリズムを適用し，単語の区切りの誤りを修正する．

1. 単語 W_i と単語 W_j に対し，2つの単語をつなぎ合わせた語 W_iW_j，W_jW_i の全文書における出現回数が多い方を合成候補Strとする．
2. 全文書における W_i，W_j 及び Str の出現回数を，NUM_{W_i}，NUM_{W_i}，NUM_{Str} とし，(1) 式により定義される値 WM を求める．WM が基準値以上ならば W_i，W_j の代わりにStrを用いる．

$$WM = \frac{NUM_{Str}}{max(NUM_{W_i}, NUM_{W_j})} \quad (1)$$

WM の基準値を高めの値をとることによって，合成後の単語がほぼ単独で出現する場合のみ，合成することができる．

4 重要語の決定

概要の手順4の詳細を述べる．本手法は，類似したクラスタを作らないように重要語を選ぶことがポイントである．よって，抽出した重要語どうしの意味的な類似度が低くなるのが望ましい．単語間の意味の類似度は，単語ベクトル V_i を用いて式（2）で計算する．単語ベクトル V_i は，単語 W_i の文書 k に対する出現頻度を要素 V_{ik} とするベクトルである．

$$cos(V_i, V_j) = \frac{\sum_{k=1}^{n} V_{ik}V_{jk}}{\sqrt{\sum_{k=1}^{n} V_{ik}^2}\sqrt{\sum_{k=1}^{n} V_{jk}^2}} \quad (2)$$

重要語を抽出する方法として，4.2節で述べる単語の重み付け手法で単語を重み付けし，以下の手順を指定回数，もしくは取り出す単語がなくなるまで繰り返す．

1. 概要の手順3で抽出した単語群 S から，一番重みの高い単語を取り出して重要語とする．
2. 取り出した単語との類似度が，基準値 C 以上の単語を単語群 S からすべて取り除く．

基準値 C は，0.05〜0.5（0.05刻み）のいずれかの値をとるものし，各値に対して実際に文書クラスタリングを行い，最も多くの種類の文書を分類できる基準値 C を採用する．なお指定した数の重要語を抽出した基準値 C は優先的に採用する．

5 本手法の拡張

本手法の拡張法として，各重要語に対して以下の手順で作成した単語クラスタを基にして，文書クラスタを作成する方法を述べる．

1. 重要語との類似度が基準値 C 以上の単語の平均類似度を C' とする．
2. 重要語との類似度が平均類似度 C' 以上の単語を単語クラスタに含める．なお複数の単語クラスタに含まれる単語は，どの単語クラスタにも含まれないものとする．

[1]電気通信大学大学院電気通信研究科システム工学専攻
[2]電気通信大学電気通信学部システム工学科
[3]http://chasen-legacy.sourceforge.jp/

本手法及び既存手法 [1,2] では「重要語を含んでいる文書集合」を文書クラスタとしたが，拡張法では「単語クラスタに含まれる単語のいずれかを含んでいる文書集合」を集合クラスタとすることによって検索結果を分類する．

6 評価

6.1 評価方法

Google の検索結果を人手により分類したものを正解データとし，本研究の手法による結果と論文 [1,2] の手法による結果を比較評価する．正解データ作成にあたり 20 代の男女 10 人に協力を頼んだ．協力者が好きなクエリを入力して，Google の検索結果 30 件をタイトルとスニペットのみから分類してもらい，15 個の正解データを得た．また情報が少ないもしくは内容があいまいで分類できないと協力者が判断した文書，及びクラスタに含まれる文書数が 1 つのクラスタは「その他」クラスタに分類し，評価には用いなかった．

本手法によるクラスタリングにおいて，3.1 節の WM の基準値を 0.6，重要語の数は各正解データの正解クラスタの数とした．なお重要語が指定した数だけ抽出できなかった場合，空のクラスタを付けて評価を行った．システムが出力したクラスタ（以下システムクラスタと呼ぶ）の評価基準として再現率（R），適合率（P），F 値 [4]，CR（Clustering Ratio）[2] を用いる．システムクラスタを SC，正解クラスタを AC としたときの R，P，F 値の計算方法を以下に示す．

$$\text{再現率}\ (R_{ij}) = \frac{|\ SC_i\text{の文書集合} \cap AC_j\text{の文書集合}\ |}{AC_j\text{の文書数}}$$

$$\text{適合率}\ (P_{ij}) = \frac{|\ SC_i\text{の文書集合} \cap AC_j\text{の文書集合}\ |}{SC_i\text{の文書数}}$$

$$F_{ij}\text{値} = \frac{2 \times R_{ij} \times P_{ij}}{R_{ij} + P_{ij}}$$

次にクラスタ集合の評価値を求めるために，システムクラスタと正解クラスタの対応付けを行う．これは「対応付けしていないクラスタの中で最大の F_{ij} となるクラスタ対を対応付けする」という処理を，全ての正解クラスタが対応付けされるまで繰り返す．

決定したクラスタ対の R_{ij}，P_{ij}，F_{ij} を R_j，P_j，F_j とし，AC_j に含まれている文書数を AN_j として (3) 式で重みづけ平均した F 値を，同様にして R，P 値を求める．

$$F = \frac{\sum_{j=1}^{n} F_j \times AN_j}{\sum_{j=1}^{n} AN_j} \tag{3}$$

また CR は，文書集合をどれだけクラスタに分類できるかの程度を表しており，式（4）で計算される．

$$CR = \frac{|\ \text{総文書数} - \text{「その他」クラスタの文書数}\ |}{\text{総文書数}} \tag{4}$$

6.2 比較手法

比較する論文 [1,2] の手法は単語の重み上位順で重要語を抽出し，本手法は 3.2 節で説明したアルゴリズムを用いて重要語を抽出する．単語の重み付け方法は論文 [1,2,6] で使われた方法を用いた．以下にその方法の概要を述べる．

- df 値 [1,2]「多くの文書に出現する単語が重要」
- tf 値 [2]「多く出現する単語が重要」

$$tf(W_i) = \sum_{j=1}^{N} tf(d_j, W_i)$$

$$tf(d_j, W_i) = \text{文書 } d_j\text{における単語 } W_i\text{の出現数}$$

- tfidf 値 [2]「特定の文書によく出現する単語が重要」

$$tfidf(W_i) = tf(W_i) \times idf(W_i)$$

$$idf(W_i) = log_2(\frac{N}{df(W_i)}) + 1$$

- SP 値 [2]「検索結果ランキング上位の単語が重要」

$$SP(W_i) = \sum_{j=1}^{N}[tf(d_j, W_i) \times sin(\frac{\pi}{1+\sqrt{j}})] \times idf(W_i)$$

d_j：ランキング j 番目の文書（下記の LP 値も同じ）

- LP 値 [2]「検索結果ランキング上位の単語が重要」

$$LP(W_i) = \sum_{j=1}^{N}[tf(d_j, W_i) \times log_N(\frac{N}{j})] \times idf(W_i)$$

- KLD 値 [6]「他の単語の出現に影響を与える単語が重要」

$$\begin{aligned} KLD(W_i) &= \sum_{j=1}^{N} kld(W_j, W_i) \\ kld(W_j, W_i) &= P(W_j|W_i) \times log\frac{P(W_j|W_i)}{P(W_j)} \\ &+ P(\neg W_j|W_i) \times log\frac{P(\neg W_j|W_i)}{P(\neg W_j)} \end{aligned}$$

$P(W_i)$：単語 W_i が出現する確率
$P(W_j|W_i)$：単語 W_i が出現する文章に単語 W_j が出現する確率
$P(\neg W_j|W_i)$：単語 W_i が出現する文章に単語 W_j が出現しない確率

- MI 値 [6]「他の単語の出現に影響を与える単語が重要」

$$\begin{aligned} MI(W_i) &= \sum_{j=1}^{N}[P(W_i) \times kld(W_j, W_i) \\ &+ P(\neg W_i) \times kld(W_j, \neg W_i)] \end{aligned}$$

7 結果及び考察

既存手法 [1,2]，本手法，本手法の拡張法における各評価基準の平均値及び平均クラスタ内文書数（Size）を図 1〜図 5 に示す．なお正解クラスタの Size は 9.04 であった．

クラスタリングの結果全体の評価指標である F 値について，各重み付け手法において既存手法より本手法の方が高い値となった．したがって，「重み上位順で重要語を抽出するより単語間の意味の関係を考慮した方が，人手に近いクラスタリングをするのに有効である」と言える．

次に CR と Size に着目する．CR も各重み付け手法において既存手法よりも本手法の方が高い値となっている．CR が高くなる理由として一般的に，類似したクラスタをあまり出力していない，もしくは Size が高いことが挙げられる．しかし本手法の Size は，正解クラスタの Size（9.04）よりも低いことから，それほど高くないことがわかる．よって，類似したクラスタがあまり作成されていないと言える．このことから，「単語間の意味の関係を考慮した重要語を選ぶことによって，類似したクラスタを極力出力せずに，多くの文書を分類できる」という考えを実現するのに適した方法であると言える．

本手法の拡張法に着目すると，F 値，CR のすべての重み付け手法において，本手法より高い値を示していることがわかる．よって重要語のみを基にクラスタリングを行うより，単語クラスタを用いた方がより人手に近い分類ができると言える．

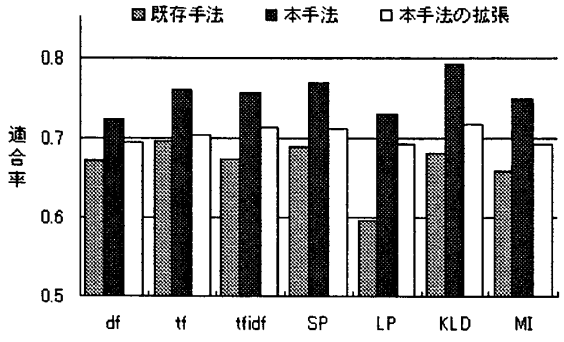

図 1: 適合率

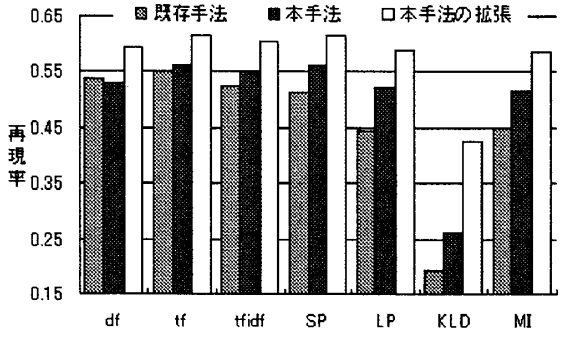

図 2: 再現率

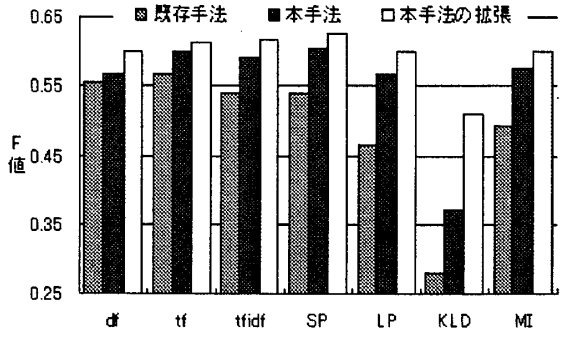

図 3: F 値

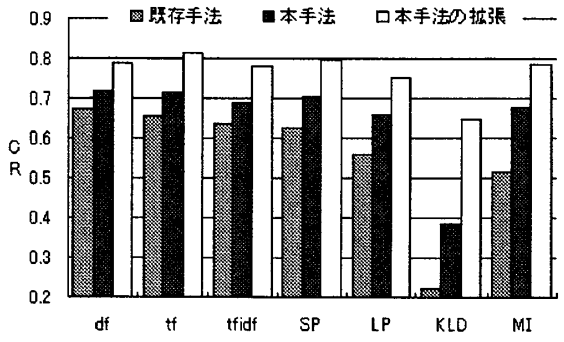

図 4: CR

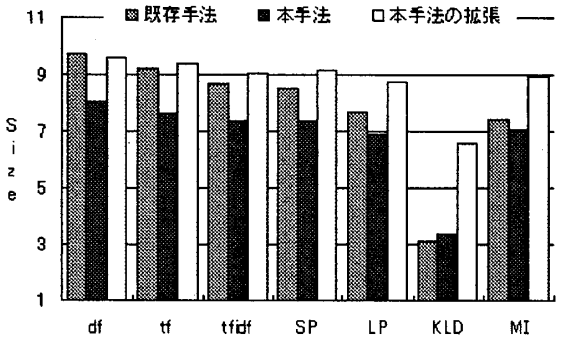

図 5: 平均クラスタ内文書数（Size）

8 おわりに

本研究では単語間の類似度を考慮して重要語を決定し，それらの重要語に基づいてクラスタリングを行う手法を提案した．評価実験を通して，本手法は単語の重み順に取り出した重要語によるクラスタリングと比べ，人手に近いクラスタリングを行うのに有効な手段であり，かつ類似したクラスタを作成しないことを示した．また，単語クラスタを用いてクラスタリングを行うことにより，より人手に近いクラスタリングを行えることも示した．

今後の課題として，拡張法の改良が挙げられる．例えば，本研究では「単語クラスタに含まれる単語のいずれかを含む文書集合」を文書クラスタとしたが，「単語クラスタに含まれる単語を n 単語以上含んでいる」，「今回の条件＋他の単語クラスタに含まれる単語を含んでいない」など文書クラスタの生成条件を厳しくすることによって，F 値の向上を試みる．また重要語と類似度の高い単語を用いて，クラスタに振り分けた後の文書の再ランキング，クラスタ内の文書の複数文書要約などへの検索支援方法の拡張が挙げられる．

参考文献

[1] 平尾一樹，竹内孔一 (2006)．複合名詞に着目した Web 検索結果のクラスタリング，情報処理学会研究報告 (2006-FI-84 2006-NL-175)，pp.35-42.

[2] 成田宏和，太田学，片山薫，石川博 (2003)．Web 文書の非排他的クラスタリング手法及びその評価方法，DB-Web2003，pp.85-92.

[3] 小熊淳一，内海彰 (2005)．語の共起情報を用いた文書クラスタリング，人工知能学会第 19 回全国大会論文集，2E1-01.

[4] 折原大，内海彰 (2007)．HTML タグの木構造に着目した Web ページのクラスタリング手法，人工知能学会第 21 回全国大会論文集，1G3-1.

[5] 大野成義，渡辺匡，片山薫，石川博，太田学 (2005)．MaxFlow アルゴリズムを用いた Web ページのクラスタリング方法の提案，データベース学会 Letters，Vol.4，No.2，pp.13-16.

[6] 若木裕美，正田備也，高須淳宏，安達淳 (2006)．検索語の曖昧性解消のためのトピック指向単語抽出および単語クラスタリング，情報処理学会論文誌データベース（TOD），Vol.49，No.SIG19（TOD32），pp.72-85.