3P-5

潜在トピックに基づく差分進化を用いた組合せ最適化による複数文書要約

重松遥 † 小林一郎 †
† お茶の水女子大学大学院 人間文化創成科学研究科 理学専攻

1 はじめに

自動要約は，大量の文書データを効率よく把握する手段として盛んに研究されており，文の組合せ最適化による手法が多く提案されている．これらの手法の多くは最適化手法として整数計画法などの厳密解法を使用しているが，厳密解法には解の探索空間が大きいほど解を求めるコストが大きくなり NP 困難となる問題がある．そこで，本研究では，厳密解を求めるのではなく，実時間の中で近似解を求める手法として効率の良い収束性が示されている差分進化 (DE)[1] を用いた要約生成を行う．また，要約として適した文の組合せとは，文書群を構成しているいくつかの話題を的確に捉えたものだと仮定し，潜在的ディリクレ配分法 (LDA)[2] により抽出した潜在トピックを用いて最適化問題の目的関数を設定する．

2 差分進化

差分進化 (DE) は個体群ベースの最適化手法の一種であり，各個体に対して次世代候補個体を生成し，候補個体の最適度が個体を上回る場合のみ置き換えていくことで，優れた個体を見つける方法である．次世代候補個体は個体と突然変異ベクトルの交叉によって生成する．ここで，DE の特徴は突然変異ベクトルにガウス突然変異ではなく，個体群の差分を利用するところにある．個体群にまとまりがないうちは差分ベクトルが大きく取られ，次世代個体群の解の幅が広がる．反対に，個体群が収束してくると差分ベクトルは小さくなるため，次世代個体群の解も収束していく．このように，突然変異ベクトルの生成に個体群の差分を用いることで，個体分布の情報が次世代個体分布に反映されるため，DE は広域探索と局所探索のバランスが自動的に取れる収束性の高いアルゴリズムと言える．決められた世代数の中で最適化を行うため，得られる解は近似解となるが，アルゴリズムの容易さ，計算速度の高速性，計算精度の高さから，最適化問題において有力な手法として注目されている．

3 潜在的ディリクレ配分法

潜在的ディリクレ配分法 (LDA) は，文書はいくつかの話題 (トピック) が混合されて作られているという仮定の下，そのトピックの確率分布を導きだすトピックモデルである．各トピック t は単語分布ベクトル ϕ_t で表され，各文書 d はトピック分布ベクトル θ_d で表される．ベクトル ϕ_t において高い確率が割り振られた単語ほど，そのトピックの特徴を表す単語となり，ベクトル θ_d によって，文書の中にどのような比率でトピックが含まれているのかを推定することができる．

4 提案手法

まず，N 個の文書からなる文書セット $D = \{d_1, d_2, ..., d_N\}$ の各文書に対して文分割を行い，n 個の文の集合として $D = \{s_1, s_2, ..., s_n\}$ のように捉えなおす．そして，n 個の文を組合わせることで要約を生成する．文の組合せは二値ベクトル $X = [x_1, x_2, ..., x_n]$ のように表現し，x_i は，文 s_i が組合せ X に含まれるとき 1，そうでないとき 0 となる．

差分進化においては，個体が文の組合せ X を表す．(ただし，DE では個体は変数ベクトルとして表されるため，0 以下の変数は 0，0 より大きい変数は 1，と二値化して考える．) 予め設定した世代数の中で要約として最適になるように文の組合せを進化させていくことで，最適度の高い組合せを見つける．最適度は適合度関数 (目的関数) によって求める．本研究では，文書は k 個の話題が集まって構成されているものと考え，なるべく各話題 t の代表文 O_t に類似しており，なおかつ，選ばれた文同士の類似度は小さくなるような文の組合せほど最適度が高くなるように適合度関数 f を設定する．

$$f(X) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\max_{t=1,2,\cdots,k}\{sim(s_{ti}, O_t)\} + \max_{t=1,2,\cdots,k}\{sim(s_{tj}, O_t)\}\right)\cdot x_i \cdot x_j}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} sim(s_i, s_j)\cdot x_i \cdot x_j}$$

$\max_{t=1,2,\cdots,k}\{sim(s_{ti}, O_t)\}$ は，文 s_i と一番近い代表文とのコサイン類似度で，文 i の重要度と考える．$sim(s_i, s_j)$ は，文 s_i と文 s_j の冗長度を表し，tf-isf 値のコサイン類似度で求める．ここで，代表文 O_t には，LDA によって推定された各トピック t の単語分布ベクトル ϕ_t を当てはめる．

Multi-Document Summarization based on latent topics using Differential Evolution

† Haruka SHIGEMATSU(shigematsu.haruka@is.ocha.ac.jp)
†† Ichiro KOBAYASHI(koba@is.ocha.ac.jp)
Advanced Sciences, Graduated School of Humanities and Sciences, Ochanomizu University (†)
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan

個体 X_i の次世代候補個体 Z_i は，個体 X_i に突然変異個体 Y_i を交叉率 CR で交叉して生成する．突然変異個体 Y_i は以下の式で求める．

$$Y_i = X_i + F \cdot (X_{best} - X_a) + F \cdot (X_{best} - X_b)$$

X_{best} は個体群の中で最も適合度が大きい個体 (ベスト個体)，X_a と X_b は個体群からランダムに選んだ 2 個体を表す．個体 X_i にベスト個体とランダム個体の重み付き差分を足す事で，ベスト個体の情報を取り入れた突然変異個体を作る．

次世代個体は，各個体と次世代候補個体を比較し，より要約に適した方を選ぶ．ここで，要約生成には要約長の制約があるため，選択の際に適合度だけでなく要約長も考慮することで制約を加味した最適化を行う．選択のルールは，(i) どちらも制約を満たす場合，適合度が大きい方，(ii) どちらかが制約を満たさない場合，制約を満たす方，(iii) どちらも制約を満たさない場合，制約の逸脱度が低い方，を選択することにする．

5 実験

5.1 実験仕様

要約評価ワークショップ DUC'04 の Task2 で使用されたデータセットを用いて提案手法 ($TopicDE$) を評価する．データセットには，話題の異なる 50 の文書セットが用意されており，1 文書セットあたり 10 個のニュース記事から成っている．各文書セットに対して，長さ 665 バイト以内の要約を生成し，評価指標 ROUGE を用いて要約の精度を測る．評価はストップワードを含めた ROUGE-1 値 "with" とストップワードを除いた ROUGE-1 値 "without" を求めた．要約は各文書セットあたり 20 回ずつ生成し，20 個のうち，最も適合度が高い個体 $TopicDE_{best}$，最も低い個体 $TopicDE_{worst}$，20 個の平均 $TopicDE_{ave}$ の ROUGE-1 値を求める．

5.2 差分進化の設定

差分進化は，10000 世代目のベスト個体を生成要約とし，個体数，交叉率 CR，差分の重み F は経験的に 50，0.7，0.45 と設定する．初期個体は通常一様分布に従った確率で数値を設定することが多く，各要素は $x = 10 - 20 \cdot rand, (0 \leq rand \leq 1)$ という式により-10 ～10 の値を一様な確率で選ぶように設定した．しかし，要約生成においては，一様な確率で要素を設定すると初期個体が 665byte を大きく上回る文の組合せになってしまう．すると，いくら進化させても制約を満たす個体が現れず，適合度を考慮できないまま世代が終了してしまうという問題がみられた (図 1 左)．そこで，予め初期個体が 665byte を下回るように，式を $x = 10 - 20 \cdot (1 - rand)^{1/6}$ と変更し，発生確率を-10 側に偏らせることで問題を解決した (図 1 右)．

(左軸：byte 数，右軸：適合度)

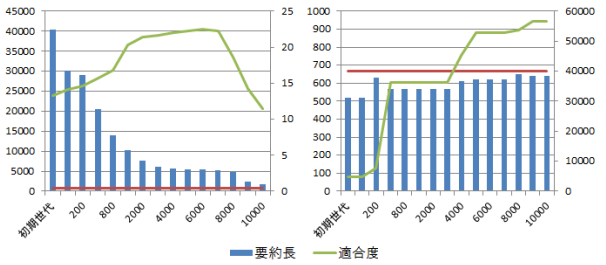

図 1: 初期個体群の操作による収束性の変化

5.3 結果と考察

50 文書セットの平均 ROUGE-1 値を表 1 に示す．$TopicDE_{best}$，$TopicDE_{ave}$，$TopicDE_{worst}$ は ROUGE-1 値の分散が小さく，差分進化によって安定した近似解を求められていることが分かった．また，$TopicDE_{best}$ が $TopicDE_{worst}$ よりも精度が高いことより，適合度の高さが要約の精度に関係していることが分かる．しかし，$TopicDE$ は，DUC'04 において ROUGE-1 値が最も高かった手法である CLASSY と比較すると，半分の精度も出せておらず，適合度関数が不十分であったと推測できる．現在，組合せの重要度を組合せの冗長度で割ることで適合度を求めているが，これだと重要度と冗長度のバランスを考慮できず，あまり重要ではないが冗長性は低い文の組合せを高く評価してしまう問題が出てくる．この問題によって $TopicDE$ の正確な精度が得られないと推測し，今後は適合度関数の再設定を課題とする．

表 1: ROUGE-1 値の評価

method	with	without
$TopicDE_{best}$	0.284	0.150
$TopicDE_{ave}$	0.283	0.142
$TopicDE_{worst}$	0.281	0.138
$CLASSY$	0.382	0.309

6 おわりに

最適化手法に差分進化を用いた要約手法を提案した．実験によって要約における差分進化の有効性を示すことができたが，適合度関数が不十分であったため，正確な精度が得られなかった．今後は，適合度関数を重要度と冗長性のバランスを考えて定義しなおすこと，他の最適化手法との比較を課題とする．

参考文献

[1] Storn R, Price K: Minimizing the Real Functions of the ICEC96 Contest by Differential Evolution, in Proc. of the Intemational Conference on Evolutionary Computation, pp. 842-844, 1996.

[2] David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research 3, pp. 993-1022, 2003.