多目的遺伝的アルゴリズムを用いた複数文書要約に関する一考察

小倉由佳里　　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情報科学コース
{ogura.yukari, koba}@is.ocha.ac.jp

1　はじめに

近年，自動要約技術の必要性が高まり，様々な手法が提案されている．自動要約の代表的な手法として重要文抽出によるものがあり，要約における重要文抽出は，最適化問題に帰着させることができる．本研究では，重要文抽出において，遺伝的アルゴリズムによる多目的最適化手法を用いる．要約の生成においては，文の結束性や冗長性，内容の網羅性，重要度等，同時に考慮しなければならない要因が複数存在する．そのため，これらを目的関数として導入し，その多目的最適化を遺伝的アルゴリズムによる多目的最適化手法である NSGAII[1] を用いて，複数の条件を満たす文の組み合わせを要約として出力する複数文書要約手法を提案する．

2　関連研究

要約生成には，重要文抽出に基づく手法を採用する研究が多くなされており，重要文抽出に最適化手法が多く適用されている．高村ら [9] は，文書の内容をより含意するような文の組み合わせを最適な要約と定義し，整数計画法を利用することで要約を生成した．また Huang ら [2] は，要約生成を多目的最適化問題と見なし，情報の網羅性，重要度，冗長性，文の結束性を定式化し，これらを目的関数として導入している．一方で Nandhini ら [3] は，文の組み合わせ最適化において，遺伝的アルゴリズムを用いて，文の結束性を考慮し生成された要約の可読性を高める手法を提案した．可読性に関連する素性として，文の長さの平均値，トリガーワードの割合，音節等を含めることにより，重要文抽出による可読性の高い要約生成を行っている．

本研究においては，文の組み合わせ最適化において，遺伝的アルゴリズムによる多目的最適化手法を用いる．文の結束性，文書のトピックと関連する度合，冗長性削減に焦点を当て，これらを考慮した重要文抽出を行う．

3　多目的遺伝的アルゴリズムによる要約生成

良い要約を生成するためには，要約長の制限や文の結束性，内容の網羅性，冗長性等のトレードオフな関係の複数の要因を同時に考慮することが求められる．これらの要因を考慮するため，要約生成における重要文抽出を多目的最適化問題と見なし，要約生成において重要であると考えられる複数の要因の定式化を行い，最適な文の組み合わせを選択する．多目的最適化には，Deb[1] らによって開発された遺伝的アルゴリズムによる多目的最適化手法である NSGAII を用いる．この手法による要約生成のアルゴリズムを以下に示す．

step 1. 初期集団の生成

遺伝子の数は，対象である複数文書の総文数とする．[0,1] の整数値をランダムに発生させ，1 つの遺伝子の中に代入する．i 番目の遺伝子は i 番目の文に対応し，これを文 s_i とする．i 番目の遺伝子が 1 である場合，文 s_i は要約に含まれ，0 である場合は，要約に含まれないことを示す．またこの時，生成された要約の要約長が制約を満たす個体のみを生成する．これにより，解が安定して収束しやすくなること，要約長の制約を満たす個体が得られやすくなることが考えられる．この個体を 50 個生成し，P_t とする（この時，$t = 0$）．

step 2. 適応度の計算

設定した 2 つの適応度関数に従って，50 個全ての個体の適応度を計算する．適応度関数については，4 章にて詳述する．

step 3. ランク付け

個体群をランク毎に分類．以下にランク付けのアルゴリズムを示す．

step i. 各個体に対して，支配されている個体の数を数える．

step ii. 支配されている個体が 0 である個体をランク r とする（初期値は $r = 1$ とする）.

step iii. step ii. でランク付けされた個体を除く.

step iv. $r = r + 1$ として step i. へ戻る．step i. 〜step iv を全ての個体がランク付けされるまで繰り返す.

step 4. 混雑度計算

個体群に混雑度をそれぞれ与える．以下に混雑度計算のアルゴリズムを示す.

step i. ランクが r である個体を適応度の値が悪い順にソートする（初期値は $r = 1$ とする）.

step ii. 適応度が最大と最小のそれぞれの個体に混雑度として無限の値を与える.

step iii. step ii. で値が与えられた個体を除いた残りの個体に対して以下の式で混雑度を与える.

$$d_i = \sum_{m=1}^{M} \frac{f_m^{i+1} - f_m^{i-1}}{f_m^{max} - f_m^{min}} \qquad (1)$$

ここで d_i は，ランク r の中でソートした個体の i 番目の個体，m は適応度の番号，f_m^i は i 番目の個体の m 番目の適応度の値である.

step iv. $r = r + 1$ とし，step i へ戻る．全ての個体に混雑度が与えられるまで step i.〜step iv. を繰り返す.

step 5. 新たな子母集団 Q_t を生成

親母集団 P_t を基に，混雑度トーナメント選択，交叉率 1.0 で交叉，突然変異率 0.1 で突然変異を行い，個体数 50 の新たな子母集団 Q_t を生成する．要約生成においては，交叉や突然変異により親個体と子個体で遺伝子の構成が大きく変化することはあまり好ましくない．なぜならば，良い親同士の交叉であっても，適応度関数の評価の低い子個体が多数生成されることが考えられるからである．そこで本研究では交叉，突然変異は Qazvinian ら [4] の手法を参考に以下のように行う.

交叉

まず一点交叉を行う．図 1(a)(b) は交叉の例である．染色体の白い部分が入れ替えられ，子個体が生成される．交叉後に親個体と子個体の 1 の数を比較して，子個体における 1 の数が親個体と異なる場合，子個体での 1 の数が親個体のそれと等しくなるよう調整を行う（図 1(c)）．調整を行う理由は，親個体と子個体で 1 の数を等しくすることにより，要約長の制約を満たす個体を生成するためである．要約長の制約を満たす親個体の文の組み合わせと，子個体のそれが類似していれば，その子個体も要約長の制約を満たすという仮定の下において行う.

突然変異

突然変異では，個体の遺伝子座において 1 と 0 が隣り合って存在している箇所を見つけ，それらの場所を入れ替えることにより行う（図 1(d)）．個体数 50 の新たな子母集団 Q_t に対して step 2. を実行する.

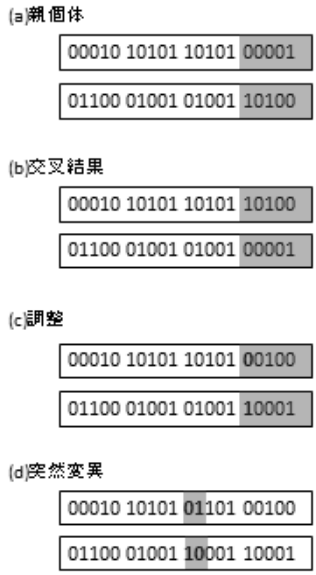

図 1: 交叉，調整，突然変異の例

step 6. $R_t = P_t \cup Q_t$ を生成

親母集団 P_t と子母集団 Q_t を合わせて，個体数 100 の新たな母集団 R_t を生成する.

step 7. R_t に対して step 3. と step 4. を実行

R_t の 100 個の個体に対してランク付けと混雑度計算を行う.

step 8. 新たな親母集団 P_{t+1} を生成

r をランクとし，その初期値を $r = 1$ とする．R_t の中からランクが小さいものから順に P_{t+1} の個体数が 50 を超えない条件の下で，新しい母集団 P_{t+1} に加える．$r = r + 1$ とし，step 8. を繰り返す．P_{t+1} の個体数が 50 より大きくなる場合は，P_{t+1} に加えずに step 9. へ移動する.

step 9. P_{t+1} の個体数を 50 にする

R_t において，P_{t+1} の個体数が 50 を超える最小のランク r を持つ個体のうち，多様に広がっているものを $50 - |P_{t+1}|$ 個 P_{t+1} に加え，P_{t+1} の個体数を 50 にする．

step 10. 世代の更新または終了

step 5.〜step 9. を設定された世代数になるまで繰り返す．設定された世代数になったら終了する．設定された世代数に満たなかったら $t = t + 1$ とし step 5. へ戻る．

4 良い要約生成のための要因

良い要約生成における重要文抽出では，抽出された文同士の結束性が高く，内容を網羅しており，かつ冗長性が低い文の組み合わせを選択することが求められる．結束性や冗長性，内容の網羅性を定式化することにより，要約生成における重要文抽出は多目的最適化問題に帰着させることができる．そこで本研究では，(i) 文の結束性，(ii) トピックに関連する度合，(iii) 冗長性削減，これらを定式化し適応度関数として用いる．(i)，(iii) の要因はトレードオフな関係であるため，遺伝的アルゴリズムによる多目的最適化を用いて文の組み合わせ最適化を行う．

4.1 文の結束性

良い要約においては，隣合う文同士が互いに高い類似度で結合しており，これが可読性の向上につながると考えられる [4]．そのため，それぞれの文間の類似度が高い，文の組み合わせを抽出する必要がある．それを考慮するため，それぞれの文間類似度の平均値を適合度関数に導入する．文間類似度は *tf-isf* から，コサイン類似度 (5) を用いて計算する．*tf-isf* は文ごとに計算され，s_j は j 番目の文，t_i は i 番目の単語を示している．

$$tf_{i,j} = \frac{t_{i,j}}{\sum_{i=1}^{n} t_{k,j}} \tag{2}$$

$$isf_i = \log \frac{N}{n_i} \tag{3}$$

$tf_{i,j}$ は，j 番目の文の i 番目の単語の単語の出現頻度であり，isf_i は，i 番目の単語が出現した文数の逆数である．ここで，N は総文数であり，n_i は単語 t_i を含む文数である．以上から単語 $w_{i,j}$ の重みは，式 (4) で計算される．

$$w_{i,j} = tf_{i,j} \times isf_i \tag{4}$$

それぞれの文間のコサイン類似度は式 (5) で表される．

$$sim(s_m, s_n) = \frac{\sum_{i=1}^{t} w_{i,m} \times w_{i,n}}{\sqrt{\sum_{i=1}^{t} w_{i,m}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,n}^2}} \tag{5}$$

文間類似度の平均値を測る適合度関数は式 (6) となる．文間類似度は，要約候補として選択された隣合う文同士が，同じ文書から抽出された文である場合にのみ測る．つまり，文 s_i と文 s_j が隣合う時，2 つの文が文書 d_i から抽出された文である場合に文間類似度を測る．

$$text\_coh_s = \frac{\sum_{s_i, s_j \in S, i<j} sim_{s_i, s_j}}{|S| - 1} \tag{6}$$

ここで，S は要約候補に含まれる全ての文の集合であり，s_j は文 s_i の次に出現する文である．

また，要約に出現する単語の共起を基に結束性を測る．単語間の結束性は，共起関係に基づく相互情報量 (MI) から得ることができる．

$$word\_coh_s(S) = \frac{\sum_{t_i, t_j \in S, i \neq j} \log\big(MI(t_i, t_j)\big)}{|(t_i, t_j) \in S| \cdot max\Big(\sum_{t_i, t_j \in S, i \neq j} \log\big(MI(t_i, t_j)\big)\Big)} \tag{7}$$

$$MI(t_i, t_j) = p(t_i, t_j) \cdot \log\Big(\frac{p(t_i, t_j)}{p(t_i) \cdot p(t_j)}\Big) \tag{8}$$

ここで，$p(t_i, t_j)$ はある文における単語 t_i と t_j の共起確率であり，$p(t_i)$ はある文における単語 t_i の出現確率である．

4.2 トピックに関連する度合

良い要約はその文書のタイトルに類似した文を含んでいる [5] ということが示されている．これは，タイトルは文書のトピックを端的に表現しているためと考えられる．そこで，生成された要約に含まれる各文とタイトルとの類似度 $TopicRelationFactor(TRF)$ を測る．方法として，文書のタイトルと各文との類似度の平均値を求める．要約 s における TRF は式 (9) で表される．

$$text\_TRF_s = \frac{\sum_{s_j \in summary} \big(1 - sim(s_j, q)\big)}{|S|} \tag{9}$$

ここで $|S|$ は生成された要約 s の文数であり，q は文書のタイトルである．

4.3 冗長性削減

文の結束性やタイトルとの関連度の高い文を抽出していくと，冗長性のある要約文が生成される可能性がある．これに対し，文の含意関係を定式化した関数を目的関数に加える．高村ら [9] は，文書要約を整数計画問題として定式化をして解く際に，内容的な観点から文 s_i が文 s_j を被覆している度合いを測ることにより，要約生成において文間の含意関係を活用している．その際，Rus ら [6] が含意関係認識においてベースラインとして用いた次の量を用いている．

$$e_{ij} = \frac{|s_i \cap s_j|}{|s_i \cup s_j|} \tag{10}$$

ここで，s_i は，その文が含む単語の集合である．よって，$s_i \cap s_j$ は，文 s_i と文 s_j に共通して含まれる単語の集合を表す．文の結束性を考慮する際に，含意関係のある文の組み合わせを多く抽出することが考えられるため，冗長性削減のために e_{ij} の平均値を適合度関数に導入する．

5 おわりに

本研究では，要約生成のための重要文抽出において，良い要約生成において重要であると考えられる複数の要因を同時に考慮した上で，文の組み合わせ最適化を行うため，遺伝的アルゴリズムによる多目的最適化手法である NSGAⅡ[1] を用いる複数文書要約の提案を行った．要約生成において考慮すべき要因として，文の結束性，内容の網羅性，冗長性削減に焦点を当て，それらを定式化し適合度関数として用いた．今後の課題としては，DUC のデータを用いた実験により提案手法の有効性を確認したいと考えている．また，より可読性の高い要約生成を行うため，文の長さ [7] や文の位置 [8] を適合度関数として含めた最適化を行いたいと考えている．

参考文献

[1] Deb K., Agrawal S., Pratap A. and Meyarivan T. 2000. : A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. Lecture notes in computer science, 1917, pp. 849-858.

[2] Huang L., He Y., Wei F. and Li W. 2010. : Modeling document summarization as multi-objective optimization. In Intelligent Information Technology and Security Informatics (IITSI), 2010 Third International Symposium on pp. 382-386. IEEE.

[3] Nandhini K. and Balasundaram S. R. 2013. : Use of Genetic Algorithm for Cohesive Summary Extraction to Assist Reading Difficulties. Applied Computational Intelligence and Soft Computing, 2013.

[4] Qazvinian V., Sharif L. and Halavati R. 2008. : Summarizing text with a genetic algorithm-based sentence extraction. IJKMS, 4(2), pp. 426-444.

[5] Silla Jr C. N., Pappa G. L., Freitas A. A. and Kaestner C. A. 2004. : Automatic text summarization with genetic algorithm-based attribute selection. In Advances in Artificial Intelligence IBERAMIA 2004 , pp. 305-314. Springer Berlin Heidelberg.

[6] Rus Graesser, McCarthy and King-lp Lin. 2005. : A study on textual entailment. In 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05) , pp. 8.

[7] Kupiec J., Pedersen J. and Chen F. 1995. : A trainable document summarizer. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval , pp. 68-73.

[8] Mani I. and Bloedorn E. 1998. : Machine learning of generic and user-focused summarization. In AAAI/IAAI, pp. 821-826.

[9] 高村大也 and 奥村学. 2010. : 施設配置問題による文書要約のモデル化. 人工知能学会論文誌, 25(1), pp. 174-182.