3I3-3

多目的遺伝的アルゴリズムを用いた複数文書要約への取り組み

小倉　由佳里　　小林　一郎
Yukari Ogura　　Ichiro Kobayashi

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
Advanced Sciences, Graduate School of Humanities and Sciences, Ochanomizu University

Automatic summarization technique, which makes a summary by collecting sentences, is regarded as a problem of combinatory optimization of important sentences. In general, to make a summary, we have to consider several factors: e.g., coherence and avoiding redundancy in a generated summary, exhaustiveness of the contents of target documents, etc. We terefore employ multi-objective genetic algorithm for optimal combination of sentences, regarding these factors as multi-objective functions. In particular, we propose a method to make a summary, applying NSGAII to solving the problem under the multi-objectives.

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る．要約における重要文抽出は，文の組合せ最適化問題に帰着させることができる．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ため，これらの要因を目的関数として導入し，多目的最適化を行う．この多目的最適化において，遺伝的アルゴリズムによる多目的最適化手法であるNSGAII[Deb 00] を用いることで，複数の条件を満たす文の組合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 [高村 09] は，文書の内容をより含意するような文の組合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 [Huang 10] 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 [Nandhini 13] は，文の組合せ最適化において，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，文の組合せ最適化において，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，要約長の制約や文の結束性，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．そこで，要約生成のための重要文抽出における組合せ最適化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，最適な文の組合せを選択する．多目的最適化には，Deb ら [Deb 00] 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
遺伝子座の数は，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 つの遺伝子座に代入する．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，生成された要約候補 S 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．個体生成の手順としては，まずランダムに選択された i 番目の遺伝子座に，文 s_i が選択されたことを示す “1” を代入し，文 s_i の文字数をカウントする．この操作を選択された文の集合 S の文字数が，設定された文字数を超えるまで繰り返し，超えたらその時点で “1” が代入されていない遺伝子には全て “0” を代入する．この操作で個体を 50 個生成し，その集団を P_t とする（この時，$t = 0$）．

step 2. 適応度の計算
設定した 2 つの目的関数に従って，50 個全ての個体の適応度を計算する．目的関数については，4 章にて記述する．

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．

連絡先: 小倉由佳里，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小林研究室，〒 112-8610　東京都文京区大塚 2-1-1，03-5978-5708，ogura.yukari@is.ocha.ac.jp

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大きなペナルティの値を与える．

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．

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

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．交叉では，一点交叉を行う．要約生成においては，良い親同士の交叉であっても，目的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ため，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．そこで本研究では突然変異に関しては Qazvinian ら [Qazvinian 08] の手法を参考に以下のように行う．

突然変異
まず，突然変異を起こす個体の持つ文の組合せから要約長を測る．その個体の要約長が制約の長さを超えていた場合は，遺伝子が “1” であるものをランダムに一つ選択し，“0” に変える．この変化後であっても要約長が制約の長さを超えている場合は，制約の長さに収まるまでこの操作を繰り返す．
制約の長さを超えていない個体の場合は，この逆の操作を行う．

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 の中からランクが高いものから順に 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 を超える最高のランク 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合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) 冗長性削減，これらを定式化し目的関数として用いる．

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 [Qazvinian 08]．そのため，それぞれの文間の類似度が高い，文の組合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目的関数に導入する．文間類似度は *tf-isf* から，コサイン類似度を用いて計算する．*tf-isf* とは，文ごとに計算される単語の重要度である．文間類似度の平均値を目的関数として導入する (式 (2))．

$$text\_coh_s = \frac{\sum_{s_i, s_j \in S, i<j} sim_{s_i, s_j}}{|S| - 1} \tag{2}$$

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 sim_{s_i,s_j} は文 s_i と文 s_j のコサイン類似度である．

しかし，式 (2) の値が最小となる文の並び順を見つけることは巡回セールスマン問題に等しく，NP 困難と呼ばれる問題のクラスに属するため，実用的な時間で見つけることが困難である．そこで本研究では，文の並び順には，文 s_i の元の文書での出現位置を用いる．要約候補の複数の文のうち，元の文書での出現位置が最も早いものを 1 番目の文として，1 番目の文と類似度の高いものを 2 番目の文，というように各文間の類似度を測っていき，その平均値を測る．元の文書での出現位置が同じ文が存在した場合は，どちらを 1 番目の文にするかランダムに選択する．

4.2 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 [高村 09] 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 [Rus 05] 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．

$$e_{ij} = \frac{|s_i \cap s_j|}{|s_i \cup s_j|} \tag{3}$$

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合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

E(式 (4)) を目的関数に導入する．

$$E = \frac{\sum_{s_i,s_j \in S} e_{ij}}{|S| - 1} \tag{4}$$

5. 実験

5.1 実験設定

対象データには，評価型ワークショップ DUC2004 の Task2 で使用されたデータセットを用いる．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，評価を行う．評価指標としては，ROUGE[Lin 04] を適用する．特に，人間の評価と相関していることが示されている，ROUGE-1 値を用いる [Lin 04]．また，ストップワードを含めた値とストップワードを除いた値を求めることにし，前者を with，後者を without として示す．要約文は各文書セットあたり，それぞれ 10 回の生成を行い，この 10 個の要約に対する ROUGE 値の平均を測る．

NSGAⅡでは，初期個体数を 50，交叉率を 1.0，突然変異率を 0.5 に設定する．また世代数は 50 世代，100 世代，300 世代と変化させる．

5.2 実験結果

50 文書セットの平均 ROUGE-1 値を 1 に示す．世代数ごとに結果を比較すると，50 世代の時，最も高い精度となっている．また，50，100，300 世代の中で，世代数が一番少ない 50 世代で精度が最大になり，世代数を増やすごとに精度が下がるという結果になっている．

表 1: ROUGE-1 値の評価

世代数	with	without
50	0.2791	0.1637
100	0.2685	0.1574
300	0.2567	0.1476

5.3 考察

世代数が少ない時ほど高い精度が得られていることから，世代数を増やした場合に，出力される解が局所解である場合があることが考えられる．原因として考えられるのは，設定された文字数を満たす要約を生成するために，初期個体群の生成において強い制限を与えていることが挙げられる．この操作のために，初期個体において選択される文の数が少ない，つまり “1” が代入されている遺伝子が少ないので，疎な個体が多数生成されている可能性が高い．そのため、交叉や突然変異を起こしても個体に大きな変化が起こらず，結果として局所解に収束してしまっているのではないかと考えられる．

6. 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，文の組合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[Deb 00] 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，冗長性削減に焦点を当て，それらを定式化し目的関数として導入し，DUC2004 を用いた実験を行った．

今後の課題としては，初期個体生成における制限を弱くすることで，多様な個体が生成されるよう改善を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 [Kupiec 95] や文の位置 [Mani 98] を目的関数として含めた最適化を行うこと，他の最適化手法との比較を課題とする．

参考文献

[Deb 00] 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.

[Huang 10] 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.

[Nandhini 13] 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.

[Qazvinian 08] Qazvinian V., Sharif L. and Halavati R. 2008. : Summarizing text with a genetic algorithm-based sentence extraction. IJKMS, 4(2), pp. 426-444.

[Silla 04] 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.

[Rus 05] 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) , p. 8.

[Kupiec 95] 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.

[Mani 98] Mani I. and Bloedorn E. 1998. : Machine learning of generic and user-focused summarization. In AAAI/IAAI, pp. 821-826.

[Lin 04] LIN, Chin-Yew. 2004. : Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, pp. 74-81.

[高村 09] 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.