人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第7回)
SIG-AM-07-04

劣モジュラ最適化としての文章情報要約

Document Summarization via Submodular Optimization

河原吉伸 (Yoshinobu Kawahara)[1*]

[1] 大阪大学 (Osaka University)

Abstract: Many criteria for document summarization is known to be submodular functions, which are the discrete counterpart of convex functions. In this paper, we review the recent studies on document summarization based on submodular set-function optimization. And, we also describe some prospects related to this field.

1 はじめに

文章情報要約は，文章を構成する文（または単語）の全体から，要約に用いられる（一部の）文（または単語）を選択する問題であり，本質的に組合せ最適化問題である．この問題は従来からも，整数計画問題や最大被覆問題などの組合せ最適化として定式化され議論されてきた経緯がある [1, 2]．そして特に近年，これらを含む従来から知られる多くの文章要約の基準が，集合関数における凸関数として知られる劣モジュラ関数である事が指摘されている [6]．この事実に基づき，この組合せ的な凸構造に基づく理論的保証を持つ効率的なアルゴリズムの適用や，種々の問題依存の構造を利用した枠組みが可能となる事が報告されている [4, 5]．

本稿では，このような背景から，劣モジュラ性を用いた文章要約に関する最近の動向について概観する．さらに，これを各種の機械学習のタスクにおいて利用する方法について述べる．本稿の以下の構成は，次のようである．まず 2 では，文章要約問題の集合関数最適化としての定式化について述べる．次に 3 では，この際の評価関数における劣モジュラ性との関係について述べ，その重要性について説明する．

2 集合関数最適化としての定式化

まず本節では，文章要約問題の集合関数最適化としての定式化について述べる．なお集合関数 f は，有限の集合 ($\mathcal{V}$ とする) が与えられたとき，その各部分集合 $\mathcal{S}$ ($\subseteq \mathcal{V}$) へ実数を割り当てる関数，つまり $f : 2^{\mathcal{V}} \to \mathbb{R}$ として定義される．

要約の対象となる文章に含まれる文 (sentence) の全体を $\mathcal{V} = \{1, \ldots, N\}$ とする (N は文の数)．このとき

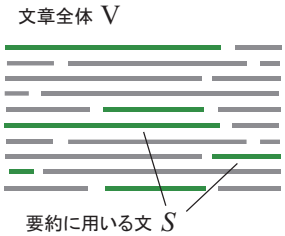

図 1: 文章要約における有限集合 $\mathcal{V}$ とその部分集合 $\mathcal{S}$ の定義の概念図．

文章要約は，要約の質をはかる基準を f を最大化にするような文の集合 $\mathcal{S} \subseteq \mathcal{V}$ を選択する問題として定式化される．このとき，要約の長さに制限があるのが一般的であるため，各文 $i(\in \mathcal{V})$ を選択する事のコストを c_i とすると，次の最適化問題が得られる．

$$\max_{\mathcal{S} \subseteq \mathcal{V}} \ f(\mathcal{S}) \quad \text{s.t.} \ \sum_{i \in \mathcal{S}} c_i \leq b \tag{1}$$

ただし，b は許容される要約の最大の長さである．

要約の質をはかる基準 f としては様々なものが提案されているが，最も一般的なものとしては，次式のように定義される被覆関数

$$\mathcal{C}_i(\mathcal{S}) = \sum_{j \in \mathcal{S}} w_{ij}$$

の和 $\sum_{i \in \mathcal{V}} \mathcal{C}_i$ が挙げられる [2]．ただし，w_{ij} は 2 つの文 i と j の類似性を表す量である．つまり $\mathcal{C}_i$ は，文 i の内容が，選択した文の集合 $\mathcal{S}$ によりどれだけ表されているかという基準になっている．これを文章全体 $\mathcal{V}$ に関して足したものは，選択した文の集合 $\mathcal{S}$ がどれだけ文章全体の内容を表すかを表す基準となる．一般には，要約 $\mathcal{S}$ が十分に文章全体を表されている場合を考慮して，次式のような基準を用いる事が多い．

$$\mathcal{L}(\mathcal{S}) = \sum_{i \in \mathcal{V}} \min\{\mathcal{C}_i(\mathcal{S}), \alpha \mathcal{C}_i(\mathcal{V})\} \tag{2}$$

*連絡先：大阪大学産業科学研究所
〒 567-0047 大阪府茨木市美穂ヶ丘 8-1
E-mail: ykawahara@sanken.osaka-u.ac.jp

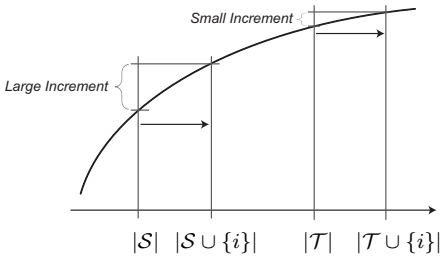

図 2: 劣モジュラ関数の定義 (3) の概念図.

これらの基準は，あとで見る劣モジュラ性 (及び，単調性) と呼ばれる性質を満たしており，効率的に良い解を得られるアルゴリズムを適用する事が可能となる．

3 劣モジュラ性の利用

劣モジュラ性は，集合関数における凸性にあたる離散構造である．上述のように，文書要約は集合関数の最適化として定式化される．その際その基準となる集合関数は，劣モジュラ性を満たす場合が多く，これにより効率的に良い解を得る事ができる．ここでは，そのようないくつかの例について述べる．

なお，人工知能分野における劣モジュラ性の利用に関しては，著者による解説 [9] なども参照されたい．

3.1 劣モジュラ関数とその最大化

劣モジュラ関数は，集合関数における凸性にあたる離散構造であり，1980 年代頃に Lovász により知られるようになった [7]．連続関数における凸性と同様，最小化が効率的に可能であり，局所最適性と大域最適性の一致性や，双対性など凸関数と類似した概念を定義する事ができる．劣モジュラ性には等価な複数の定義が存在するが，次式が直感的にも分かりやすくよく用いられる．

$$f(\mathcal{S}+i)-f(\mathcal{S}) \leq f(\mathcal{T}+i)-f(\mathcal{T}) \qquad (3)$$

ただし，$\forall \mathcal{S} \subset \mathcal{T} \subseteq \mathcal{V}, \forall i \in \mathcal{V} \setminus \mathcal{T}$ である．つまり包含関係にある 2 つの集合 $\mathcal{S}$ と $\mathcal{T}$ に関して，包含される集合 $\mathcal{S}$ へ新しい要素 i を加えた際の増分が，包含する集合 $\mathcal{T}$ の場合のそれより大きくなる (図 2 参照)．このように劣モジュラ関数は，サイズと共に増加が穏やかになる性質を持っており，限界効用逓減の法則を表す関数としても知られる．また，任意の $\mathcal{S} \subseteq \mathcal{T} (\subseteq \mathcal{V})$ に対して，集合関数 f が $f(\mathcal{S}) \leq f(\mathcal{T})$ を満たすとき，f は単調非減少であると言う．

なお上述の被覆関数は (単調非減少) 劣モジュラ関数であり，その他の基準も劣モジュラ性を満たす場合が多い [6]．これは，多くの文を使うほど，元の文章全体の内容を表せる傾向が高くなる事からも直感的に理解できる (図 3 参照).

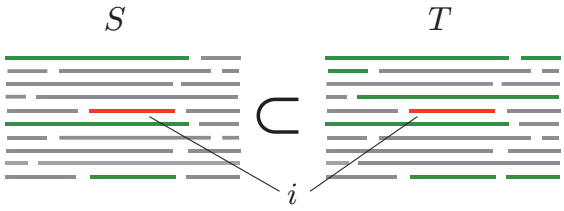

図 3: 文章要約における評価関数の劣モジュラ性 (3) の概念図.

Algorithm 1 どん欲法の手順

1: $\mathcal{S}_0 \leftarrow \emptyset, i = 1$ に設定.
2: **while** $i \leq k$ **do**
3: $\quad \rho^+_{j_i}(\mathcal{S}_{i-1}) = \arg\max_{j \in \mathcal{V} \setminus \mathcal{S}_{i-1}} \rho^+_j(\mathcal{S}_{i-1})$ となる要素 $j_i \in \mathcal{V} \setminus \mathcal{S}_{i-1}$ を選択.
4: $\quad \mathcal{S}_i \leftarrow \mathcal{S}_{i-1} \cup \{j_i\}, i \leftarrow i+1$.
5: **end while**

上述のように，文章要約によく用いられる被覆関数は最大化される事で，文章を要約する文の集合 $\mathcal{S}$ を選択する．このように文章要約は，(単調非減少) 劣モジュラ関数の最大化として定式化される事が多い (つまり，式 (1) における f が単調非減少劣モジュラ関数)．一般に，(単調非減少) 劣モジュラ関数の (サイズ制約下での) 最大化問題は NP 困難な問題であるが，どん欲法と呼ばれる単純なアルゴリズムにより，理論的に，かつ実用的に良い近似解が得られる事が知られている [8]．どん欲法の手順は，Algorithm 1 に示すように単純なものであるが，最悪ケースでも最適解の $(1-1/e) \approx 0.632$ 倍の値を持つ近似解を与える事が知られている (e は自然対数の底)．ただし $\rho^+_j(\mathcal{S}) := f(\mathcal{S} \cup \{j\}) - f(\mathcal{S}))$ である．経験的にも，多くの場合で貪欲法により極めて良い解が得られる事が報告されている [8].

3.2 要約基準と劣モジュラ性

文章情報要約のための基準は，これまで様々なものが提案されてきた．これらは一般に，選択する文の冗長性をできるだけ除外する，というのが基本的な考え方であるものが多いが，劣モジュラ性を満たす事が知られている．

例えば，一般的によく用いられる基準として，次式のように定義される (Maximal) Marginal Relevance と呼ばれる基準がある [1].

$$f(\mathcal{S}) = \sum_{i \in \mathcal{S}} [\lambda(Sim(i,Q) - (1-\lambda) \max_{j \in \mathcal{V}} Sim(i,j))]$$

ただし，$\lambda \in [0,1]$，Sim は何らかの類似尺度，Q はクエリである．この基準も劣モジュラ関数である事が示されている．また別の例としては，先の被覆関数 (2) に加え，各 $\mathcal{S}$ を選択する事に対する利得 $\mathcal{R}(\mathcal{S})$ を定義し，これらを加え合わせた基準も提案されている．

$$f(\mathcal{S}) = \mathcal{L}(\mathcal{S}) + \gamma\mathcal{R}(\mathcal{S})$$

$\mathcal{R}(\mathcal{S})$ としては，次式のように，選択する文が分散する事に対して利得を与えるようなものが知られる [6]．

$$\mathcal{R}(\mathcal{S}) = \sum_{i=1}^{K} \sqrt{\sum_{j \in P_i \cap \mathcal{S}}}$$

ただし，P_i $(i = 1, \ldots, K)$ は $\mathcal{V}$ の分割を表す．また，人による要約との比較に基づいた，ROUGE-N[3] と呼ばれる一般的に用いられる基準も，劣モジュラ性を満たしている事が知られている．

このように，要約に用いられる多くの基準は，劣モジュラ性を満たした集合関数である．従ってその最大化に関しては，劣モジュラ性のために，理論的保証のある近似解がどん欲法により効率的に得られる．また，より実用的なアルゴリズムなども多数提案されており，これらを問題に応じて適用する事で大規模な場合などでも適用可能な自動要約が可能となると言える．

4 むすび

本稿では，劣モジュラ関数最大化としての文章要約に関する定式化について述べた．劣モジュラ関数最大化は，どん欲法により効率的に理論保証のある近似解が得られる事が知られており，その他の実用的なアルゴリズムも多数提案されている．これらを適用する事により，実用的であり，かつ理論的保証のある文章要約を行う事ができる．

本稿ではふれなかったが，劣モジュラ最適化としての定式化をベースにする事により，機械学習で扱われる様々な問題と共通の枠組みの中で議論する事ができるようになる．これにより，文章と，その他のデータ (画像など) とを融合的に用いた数理的枠組みを実現する事も可能であると思われる．

参考文献

[1] J. Carbonell and J. Goldstein. The use of MNR, diversity-based reranking for reordering documents and producing summaries. In *Proc. of the Ann. Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval (SIGIR'98)*, pages 335–336, 1998.

[2] E. Filatova and V. Hatzivassiloglou. A formal model for information selection in multi-sentence text extraction. In *Proc. of the 20th Int'l Conf. on Computational Linguistics (COLING'04)*.

[3] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proc. of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004*, 2004.

[4] H. Lin and J. Bilmes. Multi-document summarization via budgeted maximization of submodular functions. In *Proc. of the 48th Ann. Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'10)*, pages 912–920, 2010.

[5] H. Lin and J. Bilmes. Word alignment via submodular maximization over matroids. In *Proc. of the 49th Ann. Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11)*, 2011.

[6] H. Lin and J.A. Bilmes. A class of submodular functions for document summa- rization. In *Proc. of the 49th Ann. Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT'11)*, pages 510–520, 2011.

[7] L. Lovász. Submodular functions and convexity. In A. Bachem, M. Grötschel, and B. Korte, editors, *Mathematical Programming – The State of the Art*, pages 235–257. 1983.

[8] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functions – I. *Mathematical Programming*, 14:265–249, 1978.

[9] 河原吉伸，永野清仁，鷲尾隆. 劣モジュラ性を用いた知能情報処理への新展開. **人工知能学会誌**, 27(3):252–260, 2012.