3I3-2

# BBS要約における整数線形計画法の適用

## Application of Integer Linear Programming to BBS Summarization

田中 駿 *1 Shun Tanaka　矢野 裕一郎 *2 Yuichiro Yano　二宮 崇 *3 Takashi Ninomiya　高村 大也 *4 Hiroya Takamura

*1愛媛大学 工学部 情報工学科 Dept. of Computer Science, Ehime University

*2*3愛媛大学 大学院理工学研究科 Graduate School of Science and Engineering, Ehime University

*4東京工業大学 精密工学研究所 Precision and Intelligence Laboratory, Tokyo Institute of Technology

We propose a summarization model for bulletin board system (BBS) based on integer linear programming. BBS summarization is one of text summarization tasks, which aims to extract important responses from BBS responses. Summarization methods based on integer linear programming are known to generate high-quality summaries in the text summarization tasks, but these methods cannot straightforwardly be applied to BBS summarization because these methods presuppose that the maximum length of summaries is given but BBS summarization does not. In this paper, we propose a method based on integer linear programming for generating variable-length summaries.

## 1. はじめに

近年，コンピュータネットワークの発達と共に，人々はインターネットを通して情報を容易に入手することが可能となった．特に，電子掲示板 (BBS) においては，誰でも話題提起やコメントを行うことができるため，多様な意見を交換する場として多くの人に利用されている．しかし，インターネット上の多くの BBS サイトでは，最大で 1,000 までの投稿を書き込むことができたり，誰でも情報を発信できたりするために，中にはトピックと関係のない投稿も存在し，トピックに関する情報のみを入手するためには手間がかかる．人手で BBS 記事を要約し，掲載している「まとめサイト」と呼ばれるウェブサイトが多数存在するが,「まとめサイト」の構築には人的，時間的コストがかかるため，BBS 記事に含まれる投稿から，必要な投稿のみを抜粋する技術の実現が期待されている．

本研究は，整数線形計画法を用いた BBS のための自動要約手法を提案する.「まとめサイト」には，ウェブサイト管理者によって BBS の要約記事が日々掲載されており，要約データを大量に入手することができる．これらのデータを機械学習し，要約を行うことが考えられるが，単純に単語ベクトルなどを入力として学習する手法では，高い精度を実現することができない．近年，文書要約 [4] を整数線形計画問題として解く自動要約手法が研究されており，非常に高い要約精度を実現している [2, 7, 8, 5]．これらの自動要約手法を BBS 要約にそのまま適用することが考えられるが，これらの手法の多くは字数制限等を整数線形計画問題の制約としており，この制約の下でより多くの情報を再現する手法となっているため，字数の制限がない BBS 要約にこれらの手法を適用することは難しい．本研究では，要約対象の記事に対して要約として採用する投稿数に関する制約と，投稿番号に関する制約を与える手法を提案する．

連絡先: 田中 駿，愛媛大学工学部情報工学科，
shun@ai.cs.ehime-u.ac.jp
(発表時は，東京工業大学 総合理工学研究科に在籍予定．)

1: 名前:A 2013/12/10 10:30:10 ID:xxxxxx 「mixiニュース」がスマホアプリに
2: 名前:B 2013/12/10 10:35:20 ID:aaaaaa いまどきmixiなんて使ってる人いるのか
3: 名前:C 2013/12/10 10:36:56 ID:bbbbbb 今なら無料 http://www.fkgame.com/
4: 名前:D 2013/12/10 11:14:10 ID:cccccc りんごたべたい
5: 名前:E 2013/12/10 11:45:50 ID:dddddd 情強はGoogleニュースだろ
6: 名前:F 2013/12/11 10:30:10 ID:eeeeee 20年間彼女いない俺はホグワーツに入学できる
7: 名前:G 2013/12/12 10:30:10 ID:ffffff おれちょっとmixiの株買ってくるわ

図 1: BBS 記事の例

## 2. BBS 要約

BBS には，たくさんの記事が存在しており，記事には 1 つの話題に対する，たくさんの人の意見が含まれている．各記事は，複数の投稿から構成されており，投稿は話題提起であるトピックか，それに対する反応であるレスに分類される (図 1)．トピックは記事中の 1 番最初の投稿であり，レスは 2 番目以降の投稿の集合である．レスの集合中には，広告等のトピックと関係のないレスが含まれる場合がある．

BBS 要約の目的は，トピックと関係のないレスを除去し，トピックに即したレスのみを抽出することである．図 1 中の黒地のレスは，トピックと関係のないレスである．このようなトピックとの関連性の低いレスを取り除き，図 1 中の投稿番号 1，2，5，7 のようなトピックに即したレスのみを抜粋した記事を生成することで BBS 要約を行う．

本研究では，人手によって BBS 記事を要約している「まとめサイト」に掲載されている投稿を要約の正解と考え，元の BBS 記事 (元記事) における各投稿に対し,「まとめサイト」において採用されていれば採用タグ，採用されていなければ不採用タグを付与し，BBS 要約データを作成する．なお,「まとめサイト」では，元記事のレス数の約 10 分の 1 程度までレスを圧縮している．

要約率をどの程度要約対象データを圧縮したかを表す指標

とし，式 (1) で定義する．

$$\text{要約率} = \frac{\text{要約採用投稿数}}{\text{元記事投稿数}} \tag{1}$$

## 3. 整数線形計画法を用いた研究

Filatova らは，整数線形計画問題として文書要約を行う手法を提案した [2]．整数線形計画法を利用するメリットとして，要約として採用された複数の文書に同じ内容を含むという冗長性へ対処することができるということが挙げられる [8]．Takamura らは，文書要約を整数線形計画問題として定式化し，字数制限や被覆，関連性の制約の下で最適な文書を選び要約を行う手法を提案した [5]．Takamura らの研究で定式化された整数線形計画法は式 (2) から式 (6) によって表される．

$$max. \quad \sum_j b_j z_j \tag{2}$$
$$s.t. \quad \sum_i x_i \leq K, \tag{3}$$
$$\sum_i a_{ij} x_i \geq z_j; \forall j, \tag{4}$$
$$x_i \in \{0,1\}; \forall i, \tag{5}$$
$$z_j \in \{0,1\}; \forall j. \tag{6}$$

$x_i$ は $i$ 番目の文が要約に採用される場合に $x_i = 1$ となる変数であり (式 (5))，また，$z_j$ は単語 $j$ が要約に採用される場合に $z_j = 1$ となる変数である (式 (6))．$b_j$ は単語 $j$ の重みを表す定数である．$a_{ij}$ は $i$ 番目の文中に含まれる単語 $j$ の数であり，$i$ 番目の文を採用する場合は単語 $j$ が $i$ 番目の文中に少なくとも一度は出現しなくてはならないという制約 (式 (4)) と，要約として採用する文の長さが $K$ 以内であるという制約 (式 (3)) の下で，要約として採用された文に含まれる単語の重みを最大化する $x_i$ と $z_j$ を求める．

## 4. 整数線形計画法を用いた BBS 要約手法

BBS 要約に適用する整数線形計画法と，整数線形計画問題の制約式に追加する投稿番号コスト制約，可変要約長制約について述べる．

### 4.1 BBS 要約のための整数線形計画法

3 節で述べた整数線形計画問題を変更し，BBS 要約のための整数線形計画問題を定義し，その整数線形計画問題を解くことにより要約を行う．式 (7) から式 (11) に，本研究で使用する整数線形計画問題を示す．

$$max. \quad \sum_j b_j z_j \tag{7}$$
$$s.t. \quad \sum_i d_i x_i \leq K(n), \tag{8}$$
$$\sum_i a_{ij} x_i \geq z_j; \forall j, \tag{9}$$
$$x_i \in \{0,1\}; \forall i, \tag{10}$$
$$z_j \in \{0,1\}; \forall j. \tag{11}$$

$x_i$ は $i$ 番目の投稿が要約に採用される場合に $x_i = 1$ となる変数であり，$z_j$，$a_{ij}$，$b_j$ は式 (2) から式 (6) で定義される変数，定数と同義である．3.1 節の整数線形計画法との差異は，式 (3) と式 (8) にある．$d_i$ は，投稿番号コスト制約によって決定される，投稿番号に対する要約採用コストである．$K(n)$ は，可変要約長制約によって決定される，要約として採用する投稿に対するコストの合計を決める関数であり，$n$ は要約対象記事の投稿数である．これらの制約については以下で詳しく述べる．また，本研究では，単語ベクトルを素性ベクトルとする L2 正則化項付ロジスティック回帰 (L2LR) を学習し，L2LR を学習した結果得られる単語に対する重みを単語に対する重み ($b_j$) とした．

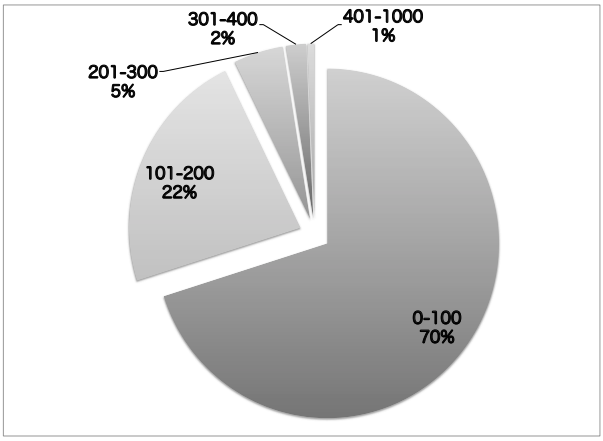


図 2: 要約データ中の投稿番号の内訳

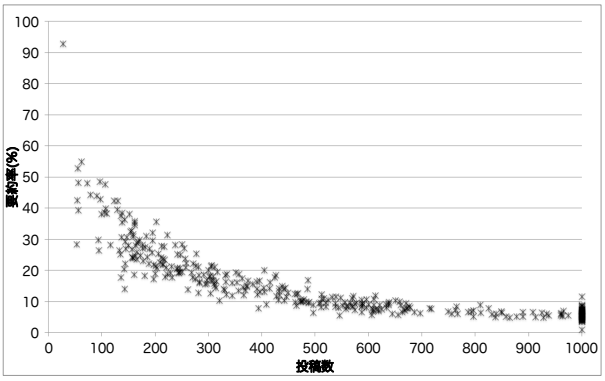


図 3: 投稿数と要約率の関係

#### 投稿番号コスト制約

図 2 は，要約として採用された投稿番号の分布である．「まとめサイト」では投稿番号の小さいレスを採用する傾向にあり，投稿番号 0-100 が 70% と，要約データの大部分を占めている．要するに，投稿番号の若い番号ほど「まとめサイト」に掲載されやすいということである．これは，たくさんの投稿が含まれる記事の場合，投稿番号が大きくなるにつれ，以前に投稿された内容と類似する投稿が出現することがあることや，「まとめサイト」の製作者が投稿番号の大きい投稿を要約対象としていないことなどが原因であると推測される．そのため本研究でも，投稿番号に応じてコストを設け，投稿番号の大きい投稿を要約として採用しないように調整する制約を設けた．投稿番号 $i$ に対するコスト $d_i$ は，式 (12) で表される．

$$d_i = e^{0.08i} + 10 \tag{12}$$

#### 可変要約長制約

本研究では，投稿数によって適切な要約長が決定されると仮定し，投稿数を $n$ としたとき，要約として採用する投稿に対するコストの合計の上限を $K(n)$ とする．図 3 は，投稿数に対する要約率のグラフである．$K(n) = \beta n$ とすると，要約率を $\beta$ で固定することになるが，例えば要約率を 10% に固定して

しまうと，投稿数の少ない記事を要約する際は，要約で採用できる投稿数がかなり少なくなってしまい，十分に要約を行うことができない．そこで本研究では，$n$ に対する 2 次式で $K(n)$ を定義する．

$$K(n) = \alpha n^2 + \beta n \tag{13}$$

式 (13) 中の $\alpha$，$\beta$ は，関数 $K(n)$ の出力値を調整するためのパラメータであり，開発データセットを使用し調整した結果，本研究では $\alpha = 0.0192$，$\beta = 24$ とした．

## 5. 実験

### 5.1 BBS 要約のプロセス

本研究で使用する，BBS 要約データの作成方法について述べる．

1. データ取得
   ウェブサイトから BBS の記事データと，要約データを取得する．正解データとして，大手掲示板サイト *i の「まとめサイト *ii」と呼ばれるウェブサイトの記事データを使用した．

2. HTML タグの除去
   取得したデータは HTML 形式なので，HTML タグの除去を行う．本研究では，改行コードである ⟨br⟩ タグは投稿中の文字の要素として扱うため，除去しない．

3. 各投稿に対し採用/不採用タグを付ける
   記事データと要約データを比較し，各投稿に対して要約として採用されている場合には 1，不採用の場合には 0 をタグ付けする．

### 5.2 実験環境

以下の 5 つの手法に対して精度の比較を行った．

1. L2LR
2. L2LR+オーバーサンプリング
3. 整数線形計画法 (要約長固定)
4. 整数線形計画法 (要約率固定)
5. 整数線形計画法 (提案手法)

L2LR は，L2 正則化項付ロジスティック回帰である．

L2LR+オーバーサンプリングは，正例と負例の割合が約 3:7 になるように要約データの正例の中からランダムに選択 (オーバーサンプリング [3, 6]) し，正例の量を補正した L2LR である．要約データ中の正例と負例の割合は約 1:9 となっており，L2LR で要約を行うと負例にバイアスがかかる結果，要約率が著しく下がってしまうため，正例の量を補正した．

整数線形計画法 (要約長固定) は，要約採用投稿数を要約対象記事の投稿数にかかわらず一意に定めたもので，本研究では開発データセットでスコアが最も高くなる要約長 $K = 850$ とする．これは，投稿番号コスト制約と可変要約長を設けない場合での提案手法と同じである．

整数線形計画法 (要約率固定) は，要約採用投稿数を要約対象記事の投稿数の 10 分の 1 とし，整数線形計画法として解いたものである．これは，投稿番号コスト制約を設けず，可変要約長制約については $\alpha = 0$，$\beta = \frac{1}{10}$ とした提案手法と同じである．

整数線形計画法 (提案手法) は，投稿番号コスト制約と可変要約長制約を加えた整数線形計画法である．

実験データとして用いる全データセットは 1,014 個の記事から成り，訓練データセットとして 812 記事，ハイパーパラメータ調整に使用する開発データセットとして 101 記事，テストデータセットとして 101 記事に分割して使用した．データセットの内訳を表 1 に示す．

素性ベクトルとして用いる単語ベクトルは 200,129 個の単語から成り，L2LR によって各単語に対する重みを学習した．投稿内の各文の単語分割には MeCab Ver. 0.994 を用い，L2LR の学習には，LIBLINEAR Ver. 1.93 [1] を使用した．整数線形計画問題の最適解を求める為の整数線形計画ソルバーとして，本実験では ILOG CPLEX Ver. 12.5.1 (IBM 社) を使用した．

本研究では，要約の精度を測る手法として f-score を使用した．f-score は Recall(再現率) と Precision(適合率) から導出され，それぞれ次式で定義される．

$$Recall = \frac{|\text{正解データ中の投稿} \cap \text{システムが出力した投稿}|}{|\text{正解データ中の投稿}|}$$

$$Precision = \frac{|\text{正解データ中の投稿} \cap \text{システムが出力した投稿}|}{|\text{システムが出力した投稿}|}$$

$$f\text{-}score = \frac{2 * Recall * Precision}{Recall + Precision}$$

表 1: データセットの内訳

| | 記事数 | 投稿数 | 要約採用投稿数 | 要約率 |
|---|---|---|---|---|
| 訓練データセット | 812 | 457,226 | 38,626 | 8.45% |
| 開発データセット | 101 | 67,596 | 4,821 | 7.13% |
| テストデータセット | 101 | 61,913 | 4,718 | 7.62% |

表 2: 実験結果

| | f-score | 要約率 |
|---|---|---|
| L2LR | 5.3% | 0.3% |
| L2LR+オーバーサンプリング | 14.5% | 10.6% |
| 整数線形計画法 (要約長固定) | 18.3% | 87.0% |
| 整数線形計画法 (要約率固定) | 6.2% | 9.9% |
| 整数線形計画法 (提案手法) | 30.2% | 11.2% |

### 5.3 実験結果

L2LR，L2LR+オーバーサンプリング，整数線形計画法 (要約長固定)，整数線形計画法 (要約率固定)，整数線形計画法 (提案手法) の各手法を用いて BBS 要約の評価実験を行った．実験結果を表 2 に示す．

L2LR での要約精度 5.3%，L2LR+オーバーサンプリングでの要約精度 14.5%，整数線形計画法 (要約長固定) での要約精度 18.3%，整数線形計画法 (要約率固定) での要約精度 6.2% と比べ，L2LR にオーバーサンプリングを用いた BBS 要約と同程度の要約率で精度が 30.2% と精度が向上した．

オーバーサンプリングによって L2LR の精度が 5.3% から 14.5% まで向上した．整数線形計画法 (要約長固定) は，整数線形計画法 (提案手法) 以外の手法と比較すると精度が向上しているが，要約率が著しく低くなっている．これは要約として採用できる投稿数 $K = 850$ と非常に多いため，Recall が非常

*i 2ちゃんねる:http://www.2ch.net/
*ii 東アジア・政治経済ニュース:http://www.m9l-o-l.com/

に高くなり，Precision は低くなったが結果的に要約精度が向上したものと思われる．また，整数線形計画法 (要約率固定) の精度が低いのは，投稿番号コスト制約がないためであると思われる．要約データでは，投稿番号の小さい投稿を多く採用しているが，整数線形計画法 (要約率固定) では投稿番号コスト制約がないため，投稿番号の大きい投稿も要約として採用する場合があった．

## 6. まとめと今後の課題

本研究では，BBS 要約に整数線形計画法を適用し，要約精度の向上を実現した．今回の実験では，人手によって BBS 記事を要約した「まとめサイト」を正解データとし，データセットを作成した．訓練データセットに対し，L2 正則化項付ロジスティック回帰 (L2LR) を学習し，学習によって得られた単語重みを利用し整数線形計画問題を作成した．そして，整数線形計画ソルバーを用いてこれらの問題を解くことにより BBS の自動要約を実現した．既存の整数線形計画法による文書要約の多くの手法は字数制限を整数線形計画問題の制約とするため，字数制限がない BBS 要約にこれらの手法を直接適用することは難しかった．本研究では，制約式中の要約長を投稿数に対する関数として与えることで BBS 要約を整数線形計画問題として定式化した．また，投稿番号の小さい投稿ほど BBS 要約に採用され易いという傾向があるため，投稿番号コスト制約を整数線形計画問題に追加した．

実験の結果，L2LR では要約精度 (f-score) が 5.3% であり，要約正解データ中の正例と負例の割合を調整する，オーバーサンプリングを用いた L2LR では要約精度が 14.5%，投稿番号コスト制約と可変要約長制約を用いず，要約長を要約前記事の投稿数にかかわらず要約長 $K = 850$ として要約を行う整数線形計画法 (要約長固定) では要約精度 18.3%，要約率を 10% に固定した整数線形計画法 (要約率固定) では，要約精度 6.2% であった．それらと比較し，提案手法ではオーバーサンプリングを用いた L2LR での要約率と同程度の要約率で，30.2% の要約精度を実現した．

これらの実験により，整数線形計画法は文書要約だけでなく，BBS 要約においても有効であることがわかった．また，BBS 要約全般において投稿番号コスト制約が有効であるとは一般には考えにくいが，本研究において「まとめサイト」を正解データとした場合には非常に有効であることがわかった．

今後の課題として，更に BBS 要約に特化した整数線形計画問題への変更，素性ベクトルとして引用関係等を採用すること，国外の電子掲示板を正解データとして実験を行い，より汎用性を高めることなどが考えられる．

## 参考文献

[1] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, 2008.

[2] E. Filatova and V. Hatzivassiloglou. A formal model for information selection in multi-sentence text extraction. In *Proc. of COLING 2004*, pp. 397–403, 2004.

[3] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Trans. on Knowl. and Data Eng.*, Vol. 21, No. 9, pp. 1263–1284, September 2009.

[4] I. Mani. *Automatic Summarization*. John Benjamins Publisher, 2001.

[5] H. Takamura and M. Okumura. Text summarization model based on maximum coverage problem and its variant. In *Proc. of EACL 2009*, pp. 781–789, 2009.

[6] G. Weiss, K. McCarthy, and B. Zabar. Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? In *DMIN*, pp. 35–41, 2007.

[7] W.-T. Yih, J. Goodman, L. Vanderwende, and H. Suzuki. Multi-document summarization by maximizing informative content-words. In *Proc. of IJCAI-07*, pp. 1776–1782, 2007.

[8] 西川仁, 平尾努, 牧野俊朗, 松尾義博, 松本裕治. 冗長性制約付きナップサック問題に基づく複数文書要約モデル. 自然言語処理, Vol. 20, No. 4, pp. 585–612, sep 2013.