量子コンピューティングによる記事の組み合わせ最適化

木内美波　(指導教員：工藤和恵)

1　はじめに

現在、パソコンやスマホなどの電子機器でニュースを見る機会が増えた。ネットニュースは日々発行されており、日本経済新聞社が一日に発行してる記事は約900 本である。ユーザーはその全ての記事に目を通すことは難しく、おすすめの記事を見逃してしまうこともある。大量の記事の中から、様々なジャンルのおすすめの記事のいくつかをピックアップしてユーザーに提供できれば、より効率的な情報の提供が可能になる。そこで、組み合わせ最適化を用いたプログラムでおすすめの記事の組み合わせを求めることを目標とし研究を開始した。組み合わせ最適化とは、大量の記事の組み合わせのような膨大な数の組み合わせの選択肢の中から最適な解を求める問題である。例えば、巡回セールスマン問題は有名な組合せ最適化問題の例である。

解きたい問題はハミルトニアン (目的関数) で表す。本研究のように考慮したい項目が複数ある時に、その全てをハミルトニアンで表現する。使用するハミルトニアンは 0 か 1 の値を取る二値関数を用いる QUBO (Quadratic Unconstrained Binary Optimization) 形式で作成する。ハミルトニアンが最小値をとるような変数の組み合わせが最適解となる。

使用したソルバーは量子アニーリングマシンの D-Wave である [1]。量子アニーリングマシンとは組み合わせ最適化問題を高速に解くことができる量子マシンで、ハミルトニアンの最小化を目的としている。プログラムの作成には組み合わせ最適化に関する先行研究を参考にした [2]。

2　設定

使用する記事のデータには発行された日時、タイトル、文字数、推薦度スコア (以下スコア)、ジャンルの情報が含まれている。この研究で求めたい最適な組み合わせとは、ジャンルの重複度と合計文字数の制約を満たし、かつできるだけ合計スコアが高い組み合わせのことを指す。つまりジャンルが被らず、合計文字数が理想に近い組み合わせの中で、さらにスコアが高い組み合わせを求めたいということである。また、おすすめの記事の組み合わせをユーザーに一日に 3 回提供する場合を考えることにした。よって記事の発行された時間が前日の 18 時から 6 時までを朝の記事、6 時から 12 時までを昼の記事、12 時から 18 時までを夕方の記事として分類し、それぞれの時間帯で最適な組み合わせを求める。

3　モデル

偏ったジャンルの記事ばかりが選ばれるのを避けるため、ジャンルの重複度にペナルティを課す。合計文字数については、理想の文字数を朝は 6000 字、昼と夕を 3000 字と設定し、この理想の文字数との誤差にペナルティを課す。ペナルティを小さく、かつスコアが高くなる組み合わせを求めるために次のようなハミルトニアンを作成した。

記事の組み合わせ最適化のモデルとして、次式のようなハミルトニアンを使用した。

$$H = w_1 H_1 + w_2 H_2 - H_3 \tag{1}$$

ここで、w_1, w_2 は正のパラメタである。値を変えることで項がどれほどハミルトニアン全体の値に影響するかが異なる。つまりパラメタの値で結果も変わると考えた。三つの項からなるハミルトニアンで、設定した重複度や文字数の制約も項の中身に組み込んでいる。

一つ目の項は文字数に関する項で、次式で与える。

$$H_1 = \Big(\frac{1}{\alpha}\frac{1}{L^*}\big(\sum_i l_i x_i - L^*\big)\Big)^2 \tag{2}$$

ここで、x_i は二値変数であり、記事 i が選択された時に $x_i = 1$ とし、選ばれなかった場合 $x_i = 0$ とする。l_i は記事 i の文字数である。L^* は設定した理想の合計文字数である。合計文字数と理想の合計文字数との差を理想の合計文字数で割り、文字数の理想との誤差を表す。さらにこの値を α で割る。本研究では誤差 10%を制約の範囲内とする。すなわち $\alpha = 0.1$ とする。よって実際の文字数の誤差がちょうど 10%の時、H_1 の値が 1 になる。

二つ目の項は重複度に関係する項で、次式で与える。

$$H_2 = \frac{1}{\beta}\sum_{i<j} f_{i,j} x_i x_j \tag{3}$$

記事 i と記事 j が同じジャンルだった場合 $f_{i,j} = 1$、そうでない場合は $f_{i,j} = 0$ とする。β は重複度の許容度を表す。本研究では重複度 3 以下を制約の範囲内とする。よって重複度がちょうど 3 のとき $H_2 = 1$ となるように $\beta = 3$ とする。

三つ目の項はスコアに関する項で、次式で与える。

$$H_3 = \sum_i \frac{p_i - p_{\min}}{p_{\max} - p_{\min}} x_i \tag{4}$$

p_i は記事 i のスコア、$p_{\min}$ はスコアの最小値、$p_{\max}$ はスコアの最大値である。スコアの最大値と最小値の差を基準にした正規化を行っている。

H_1 と H_2 は値が小さいほど、よく制約を満たすことを意味している。H_3 は値が高いほど、スコアが高いことを意味している。よって式 (1) で表したハミルトニアンが最小の値を取るとき、求めたい最適な記事の組み合わせを得られる可能性が高い。

4　方法

量子アニーリングで計算を行うために、元の記事のデータの値をいくつか調整した。ジャンルなしの記事は予め除外した。複数のジャンルに分類されている記事は、記事の特性を最もよく捉えたジャンルを用いた。つまりハミルトニアンに使用する全ての記事には必ず 1 つジャンルが与えられている。次に、6000 字以上の記事についてはハミルトニアンに組み込む際に全て 6000

字として扱うことにした。文字数が 6000 字よりも多い記事は数は少なく、例外的な存在ながらも、結果に影響を与えるものが多いためである。

D-Wave は全結合問題を解く場合に扱える変数の数に限りがある。また、試行の結果、100 以上の記事を一度に計算するのは難しいことが分かった。そのためプログラムに使用する記事をスコアが上位の記事 30, 40, 50 個の 3 パターンに絞り込んだ。

D-Wave では結果にムラがあったり、一回の実行でうまく答えが出ない場合もあるため、パラメタの値の 1 つの組に対して、その最小化を 1000 回繰り返す。その解の中で制約を満たし、かつ最も高いスコアが得られる結果を出力する。

パラメタ w_1, w_2 は $0.1 \leq w_1 \leq 0.5$, $0.1 \leq w_2 \leq 0.5$ の範囲で設定した。

5 結果

5.1 スコアの比較

古典的に計算した場合と D-Wave で計算した場合のスコアの違いを観察するために、シミュレーテッドアニーリング (以下 SA) でも同様の計算を行った。図 1 は、SA で計算した場合のスコアのヒートマップである。横軸を w_1、縦軸を w_2 とし、各点の色がカラーバーの通りにスコアを表している。データは 3 月 17 日の月曜の朝の記事で、計算にはそのうちのスコア上位の 30 個の記事を使用した。

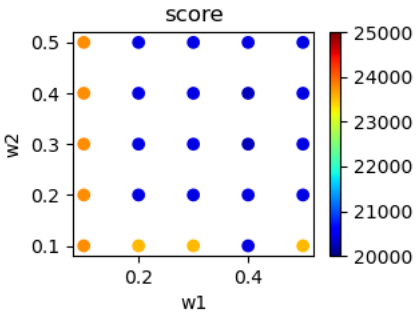

図 1: SA で計算したスコアのヒートマップ

図 2 は、D-Wave の結果で作成したヒートマップである。使用した記事や条件は図 1 で用いたものと同様である。

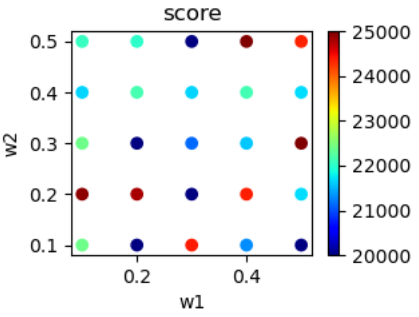

図 2: D-Wave で計算したスコアのヒートマップ

図 1 を見ると、同じ色が近くに並んでいることが分かる。一方で図 2 では、多様な色がバラバラに並んでいるように見える。また、同じパラメタの組み合わせで比較すると、D-Wave は SA より高いスコアの点もあることが分かる。よって D-Wave は SA と比べて、同じ解に固定されず計算の度に多様な解を求められる特徴があることが分かった。よって、D-Wave ではスコアが高い特定のパラメタの組み合わせを求めるよりも、パラメタの範囲を更に絞り、その中で高いスコアが出る組み合わせを都度選ぶやり方が望ましい。

5.2 記事数を変えた結果の比較

計算に使用する記事の数を変えた 3 パターンの結果を比較する。図 2 は記事数 30 個で実行した結果だが、40 個で実行した結果を図 3、50 個で実行した結果を図 4 に示す。記事数以外の条件は全て同じである。

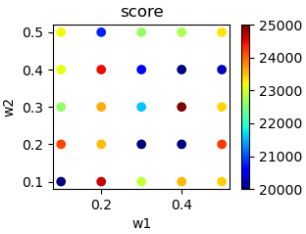

図 3: 記事数 40 個の実行結果

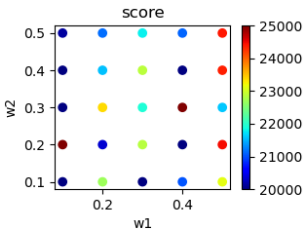

図 4: 記事数 50 個の実行結果

図 3 と図 4 はどちらも D-Wave で実行した結果であり、図 2 と同様に多様な解が求められたことが分かる。試行の結果、記事数 50 個での実行はやや精度が悪くなることが分かったが、記事数を変えたことによるスコアの違いについては明確な考察を得ることが出来なかった。最も精度が良くなる記事数を更に絞り込むことは今後の課題の一つである。

6 まとめ

本研究では、量子アニーリングを利用して最適な記事の組み合わせを求めた。D-Wave では一回のアニーリングにかかる時間は数十〜数百 μs と短い。短時間で多様な解を求めることが可能であり、その中には SA では手間と時間をかけないと得られないような高いスコアの解も含まれていた。よって高いスコアの組み合わせを効率よく求めたいという本研究の目的には、D-Wave を使用することに優位性があると言える。

今後の研究の課題として以下が挙げられる。今回は特定の日付のみで調整を行った。より一般的に使用できるようにするために違う週の同じ曜日のデータで実行する際の結果の違いを観察する必要がある。また、記事数を比較した結果についての十分な考察が出来なかったため、今後は計算に使用する記事数が変わると結果の精度に影響するのかどうかを調べ、使用する上位記事数を更に絞りこむ。そして量子アニーリングに関して、現在とは違う設定でも実行し、スコアが変化するかを試す。

参考文献

[1] D-Wave, 量子コンピューティング, https://dwavejapan.com/system/ (2021 年 12 月 17 日アクセス).

[2] N. Nishimura, K. Tanahashi, K. Suganuma, M. J. Miyama, and M. Ohzeki, Item Listing Optimization for E-Commerce Websites Based on Diversity, Front. Comput. Sci. **1**, 2 (2019)