教材推薦最適化における相転移

理学専攻・情報科学コース　岡田　翔子

1　はじめに

自分に合った教材を選び出すのは容易ではない。実際、本学の外国語教育センターのスタッフの方から、学生それぞれに合った教材を自動的に推薦してくれるようなアプリケーションがあれば便利だという声を頂いた。そこで、本研究では集合被覆問題を応用することで、学生一人一人に合った教材を推薦することを考えた。今回は、どの程度の教材数のボリュームがあれば、またどの程度の時間制約の下で学生の学習目的を満たすような教材がみつかるのか調べるために、教材がみつかる確率を成功率とし、教材数、学生の学習目的の種類、提案する教材の数、希望学習時間、それぞれを変化させた時に成功率がどう変化するのか考察した。横軸を「希望学習時間」、縦軸を「成功率」とし、教材数を変化させたところ、成功率の急激な変化が見られた。本研究では、この成功率の相転移について議論する。

2　重み付き集合被覆問題

このレコメンド問題では、「スピーキングの勉強がしたい」、「ライティングの準備がしたい」といった学生の目的と、「スピーキングの教材」、「ライティングの教材」といった教材のカテゴリーをマッチングする。同時に、例えば「1 日 30 分 3 か月程勉強したい」という学生の学習時間内に、教材の所要時間が収まるよう考慮する。これを実現するにあたり、重み付き集合被覆問題を考える。

重み付き集合被覆問題では、要素集合 $I = \{1, \cdots, m\}$ の部分集合族を $P_j (j \in N = \{1, \cdots, n\})$ とした時に、I のすべての要素をカバーする集合 P_j を幾つか選び、選んだ集合につけられた重みの総和を最小化する [1]。本研究では、学生の学習目的の集合を I とし、I を満たすような、教材のカテゴリーの集合族 P_j を求める。この時、学生の学習時間の希望を満たすために、教材の所要時間が学生が指定する学習時間の上限を超えないという条件を付与する。ここでテキスト i に要する時間を $T_t(i)$ とし、学生が学習にかけられる時間を T_u としたとき、学生の学習目的を満たすために、教材が数冊以上選ばれる可能性があることを考慮し、教材のかかる時間数の合計を T_t とすれば、時間制約は以下のように表される。

$$T_t = \sum_i T_t(i) < T_u \tag{1}$$

3　データの準備

本学の外国語教育センターの教材は 120 程度だったため、データ数を小さくしたり、大きくしたりした場合の成功率も検証するために、実データに基づいた、乱数データを作成した。まず、教材をスピーキング, ライティング, リーディング, リスニングの 4 種類のタグの組み合わせを 0 から 15 までの数値で識別できるようにした。その後、外国語教育センターの実際の英語教材のデータを教材の種類によって、先ほどの 0 から 15 までの数値に変換した。そして、タグごとのページ数の分布を調べるために、各々のタグについてヒストグラムを作成した。また、ヒストグラムでわかった各ページの頻度から、タグごとに教材の所要時間 (ページ数) の頻度の線グラフを作成し、線グラフの形が似ているもの同士をグループ分けした (図 1)。その結果、タグの数によって近似グラフの形が大きく 3 つに分かれた。これは、スピーキング、ライティングといった教材の分野に関わらず、教材の時間の分布は同じだったため、タグの数によって形が分かれたためだと考えられる。図 1 はタグが 1 種類の場合の例である。

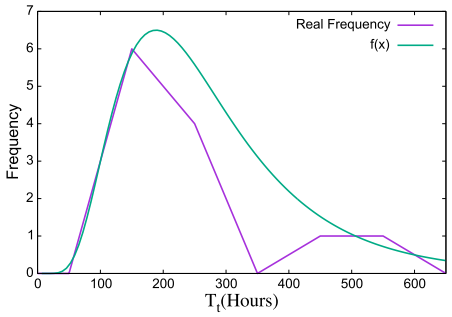

図 1: 教材の所要時間の頻度の線グラフ (タグ: 1 種類) とその近似曲線.

同時に、実データのタグの種類の度数分布表も作成した。プログラムで乱数を振った際に、タグの種類の分布がこの度数分布に従うようにすることで、データ数が小さくなったり、大きくなったりしても、実際の分布に従うようにした。

次に、図 1 の線グラフをもとに、各タグのグループごとに対数正規分布に従う近似グラフを求め、近似式を求めた (図 1)。タグの種類と同様に、プログラムの方で時間の分布がこの近似式に従うようにすることで、データ数が小さくなったり、大きくなったりしても、実際の分布に従うようにした。求めた近似式は以下の通りである。

タグが 1 種類:

$$f(x) = \frac{1785}{\sqrt{2\pi}0.51x} \exp\left[-\frac{(\log(x) - 5.5)^2}{2 \times (0.51)^2}\right] \tag{2}$$

タグが 2 または 4 種類:

$$f(x) = \frac{2500}{\sqrt{2\pi}0.54x} \exp\left[-\frac{(\log(x) - 5.58)^2}{2 \times (0.54)^2}\right] \tag{3}$$

タグが 3 種類:

$$f(x) = \frac{1010}{\sqrt{2\pi}0.3x} \exp\left[-\frac{(\log(x) - 5.76)^2}{2 \times (0.3)^2}\right] \tag{4}$$

4　方法と結果

次に、乱数で作成した教材データをもとに、教材数 N を 1000 冊から 10 万冊に変化させた際の成功率の

グラフを作成した。今回はタグの種類が 1 種類の時のグラフを示している。図 2 では、横軸が「希望学習時間」、縦軸が「成功率」となっており、プログラムを 1 万回実行した時の平均をとっている。図からわかるように、教材数を増やせば増やすほど、学生の学習目的を満たすような教材が増えることになり、短時間でも成功率が 1 に近づいている ことがわかる。教材数 N が 10 万の時に、20 時間弱で成功率が立ち上がり始め、ほぼ垂直に 1 まで到達しているのは、近似曲線 (図 1) で約 19 から立ち上がっているためだと考えられる。

10 万冊以上はデータを取るのが難しかったため、ここで、有限サイズスケーリングを行う [2]。スケーリングのパラメータを探すために T_c を成功率のちょうど真ん中である、0.5 の時の希望学習時間とし、「教材数」と「T_c」についての両対数グラフを作成した。

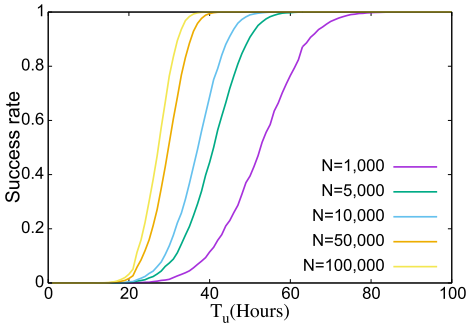

図 2: 成功率の希望学習時間依存性.

図 3 は、教材数 N が 5 から 10 万の時の T_c の両対数グラフである。両対数を作成した際に、末端がほぼ直線になっており、この直線の傾き γ は約 -0.13 となっていた。よって、$T_c \propto N^{-0.13}$ の関係があるのではないかと考えた。

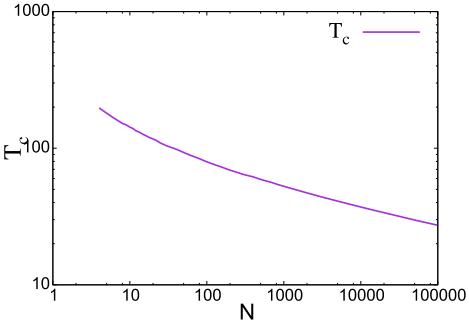

図 3: T_c の教材数依存性.

そこで、γ を用いて x 軸を $X = T_u/N^{\gamma}$ と変更し、縦軸は成功率のままでグラフをプロットしたところ、グラフが重なり合い、一点で交わった (図 4)。そこで交わった点を α とし、x 軸を $\xi = (X - \alpha)N^{1/\nu}$ と変更しさらにスケーリングし直した (図 5)。すると、5 本の線がほぼ一本に重なった。ここで ν は臨界指数であり、$\nu = 18$、$\alpha = 103$ となっていた [3]。

5 考察

実際にアプリケーションに応用にするにあたっては、レベル分けや在庫の影響などを考慮する必要がある。

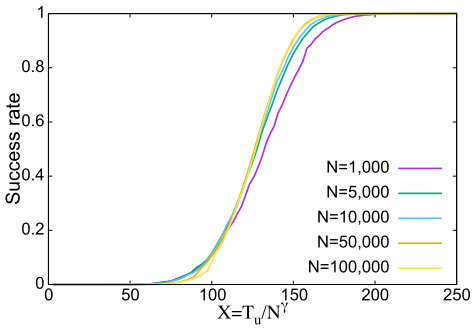

図 4: 成功率の X の依存性.

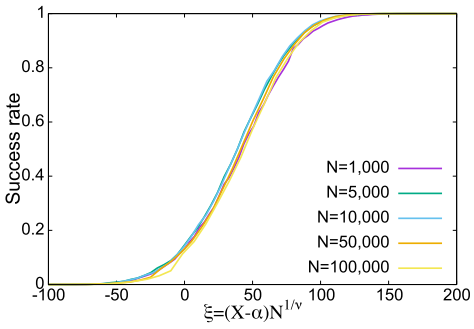

図 5: 成功率の ξ の依存性.

まず、レベル分けを考慮した場合の成功率は、各レベルの冊数においての成功率と同等であると考えれば良いので、ここでは取り上げなかった。例えば、初級の教材が 2 割ある場合、「全体の冊数 $\times$ 0.2」の冊数における成功率を考えれば良い。次に、在庫の影響については、人気教材の在庫を必要に応じて増やすことで対処できると考えられる。本研究の成功率では、教材数が増えれば増えるほど、短時間で終了できるような教材が増えることになっているが、現実世界ではそうとは限らない点にも注意しなければならない。

6 まとめ

本研究では、成功率に関する相転移を確認した。また有限サイズスケーリングをしたところ、異なるデータが一本の直線に重なり、臨界指数 ν を予測できた。また、今回はタグの種類が一種類の場合を示したが、他のタグの種類を試した場合も同様の結果が得られた。よって、これらの過程を逆変換することで教材数がより大きくなった場合の成功率を予測できる可能性があることがわかった。

参考文献

[1] M. Yagiura, M. Kishida, and T. Ibaraki. A 3-Flip Neighborhood Local Search for the Set Covering Problem, European Journal of Operations Research, **172**, 472–499, 2006.

[2] I. P. Gent, and T. Walsh. The TSP phase transition, Artificial Intelligence, **88**, 349–358, 1996

[3] 西森　秀稔，新物理学シリーズ 35—相転移・臨界現象の統計物理学，培風館，2008.