Vol.2018-CLE-26 No.4
2018/12/7

授業中の学習者のページ遷移のレーベンシュタイン距離による分析の試み

中野 裕司[1,a)]　古川 雅子[2]　大渡 拓朗[3]　久保田 真一郎[1]　杉谷 賢一[1]　島田 敬士[3]

概要：電子教科書を用いた面接授業における学習者のページ遷移に対して、レーベンシュタイン距離を用いた分析を試みた。授業中の電子テキストの操作イベントの記録データから、ページ滞在時間を無視したページ遷移のシーケンスと、そこからレーベンシュタイン距離を求めた。学生のレーベンシュタイン距離を求める対象として、担当教員のシーケンスを用いる場合と、その授業における全員のレーベンシュタイン距離の合計を最も低くする学生を選び、そのシーケンスを用いる場合について調べた。後者は、教員のシーケンスデータを必要とせず、自習や小テスト受験等、教員のシーケンスデータがない場合もレーベンシュタイン距離を求めることができる。求めたレーベンシュタイン距離に関して、クラス、担当教員、授業回、授業時間、説明時間、小テスト等に関して、その関連性を調べた。まだ予備的な分析の段階ではあるが、個々の授業回の中ではレーベンシュタイン距離と成績の間の関係は見いだせていないが、集計データに関しては、レーベンシュタイン距離を基準値 (対象となるシーケンスの長さ) の上下に分けることで、ある程度の関連性が見えたので報告する。

1. はじめに

近年、Learning Analytics (LA) が学習支援や IR (Institutional Research) において注目を集めており、学生のドロップアウトの早期発見や学習状況の詳細な把握、データに基づくフィードバック等に応用され、今後さらに発展していくと思われる [1–3]。

このような状況の中、2018 年 9 月、日本教育工学会 SIG「教育・学習支援システムの開発・実践」第 10 回研究会が開催され、その中で提供されたデータに対してチームで分析に取り組むデータハッカソンが催された。そこで我々が選んだデータは、謝辞にも述べるが、九州大学提供の電子教科書の匿名化されたログデータで、授業外も含めて、学習者のページ遷移や各種イベントが授業期間外も含めて大量に記録されているものであった。また、毎回の授業の最後に小テストも実施され、そのデータも記録されていた。この中で、我々のチームは、特に授業中の学習者の振る舞いについて着目し、学習者の残した各種イベントスの分析、k 平均法によるクラスタ分析、学習者のページ遷移 (ページシーケンス) のレーベンシュタイン距離による分析等のアプローチを試みた。

ハッカソンの限られた時間内では、なかなか具体的な結果まで到達することは難しかったが、興味ときっかけには十分なったと思う。九州大学のご厚意で、ハッカソン以降もデータの使用を許諾いただいたため、その後も分析を継続し、まだ十分な結論が得られた状況ではないが、レーベンシュタイン距離と小テストにある程度の関連性が見えてきたので、研究会で報告することとした。

レーベンシュタイン距離は、2 つの文字列がどの程度異なっているかを示すよく知られた指標で、一方の文字列に 1 文字単位の変更 (挿入、削除、置換) を最低何回行えば 2 つの文字列が一致するかといったものである [4]。LA においても、学習シーケンスは重要な分析対象であり、シーケンス間の差異を調べるのに、レーベンシュタイン距離 [7] やその他の手法 [8] が用いられる。実際のデータはページ遷移だけではなく、同一ページ内の各種イベントであったり、ページ遷移の時刻も詳細に記録されている。ここでは、ページ遷移シーケンスに単純化したデータでどの程度のことがわかるか、また単純化の効果はどのようなものか、また、軽量手法であるためリアルタイム処理への応用の可能性はなどについて、将来的には、なんらかの知見が得られないかと考えている。

[1] 熊本大学 総合情報統括センター
Center for Management of Information Technologies, Kumamoto University, Kumamoto 860–8555, Japan
[2] 国立情報学研究所 National Institute of Informatics
[3] 九州大学 Kyushu University
[a)] nakano@cc.kumamoto-u.ac.jp

表 1　コースの基本情報

コース (クラス) 名	教員	学生数
A	A	133
B1	B	120
B2	B	114
C1	C	160
C2	C	132

2. 研究方法

2.1 対象データ

使用した学習ログデータは、謝辞にも示すとおり九州大学様のご厚意によるもので [5,6]、8 週間に渡って開講された情報系科目 5 コースで収集されたもので、5 コースすべてで同じ教科書を使い、講義は座学形式で各回 90 分で実施され、表 1 に示すような構成となっている。

ここで、教員 B のみ教員のスライド操作のログも提供されているため、後述の教員のページ遷移とのレーベンシュタイン距離を利用できるのは B1,B2 コースのみである。また、C1、C2 コースに重複登録データがあったため、それらは予め両コースから外して分析を行った。

2.2 分析環境

集計、可視化、分析は、全て R の統合開発環境 RStudio を用いた。元の学習ログデータが CSV 形式であったため、全て MariaDB にテーブルとして導入し、適当にインデックスを設定し、R から RMySQL で随時読み込んで利用し、stringdist、tidyverse、GGally、ggplot2、scales、knitr 等のライブラリを適宜利用した。分析結果は R Markdown 形式で記録した。

3. 分析結果と考察

3.1 ページ遷移のシーケンスとレーベンシュタイン距離

3.1.1 担当教員のデータを使用しない場合

担当教員のページ遷移シーケンスデータがあれば、それと各学生のシーケンスデータの間でレーベンシュタイン距離を計算するのが基本的な考えであろうが、担当教員の操作データがない場合はそうもいかない。また、電子ブックを用いている授業なので、インターネット接続や LMS の利用も前提であるので、個人やグループでのインターネットを利用した学習活動や、LMS 上の小テストの利用も考えられ、その場合は担当教員のページ遷移とは殆ど無関係に、学生は電子ブックのページ遷移も行う可能性がある。

今回は、担当教員のデータを使用しない場合、その授業における全学生のレーベンシュタイン距離の合計を最も低くする学生を選び、そのシーケンスを用いた。実際には、全学生を 1 名ずつレーベンシュタイン距離の対象者に仮定し、その時の他の学生のレーベンシュタイン距離を全員分

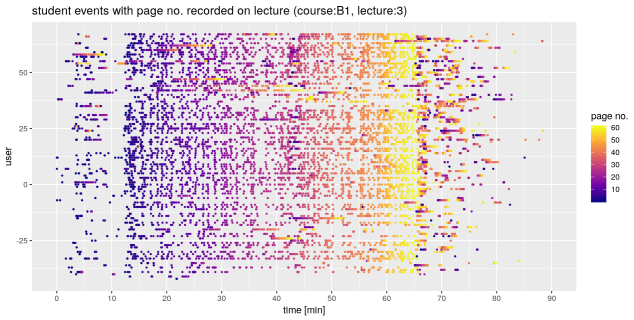


図 1　授業中に学生が行ったページ操作に関するイベント (course:B1, lecture:3)。

表 2　図 1 に関するデータ (抜粋)。

user (縦軸)	シーケンス長	レーベンシュタイン距離	ページシーケンス	シーケンス文字列
67	515	417	1, 2, 3, 2, 3, 4, 5, 4, 5, 4, ...	ぁあぃあぃいぅいぅいぃいぅうぇえぉおかが...
60	359	270	1, 2, 3, 4, 5, 6, 7, 8, 9, 10...	ぁあぃいぅうぇえぉおかがかきぎくぐくぎ...
40	276	190	1, 2, 3, 4, 5, 4, 5, 6, 7, 8, ...	ぁあぃいぅいぅうぇえぉおかがきぎくぐけげ...
20	196	141	1, 2, 3, 2, 1, 2, 3, 4, 5, 6, ...	ぁあぃあぁあぃいぅうぇえぇおおかがきぁあぃ...
2	148	83	1, 2, 3, 4, 5, 4, 5, 6, 5, 7, ...	ぁあぃいぅいぅうぅぇえうえぇおおぉおかがきぎ...
1	157	74	1, 2, 3, 4, 5, 6, 7, 8, 9, 10...	ぁあぃいぅうぇえぉおかがきぎくぐくぎきが...
0	147	0	1, 2, 3, 4, 5, 6, 7, 8, 9, 8, ...	ぁあぃいぅうぇえぉえぇえぇぉぁあぃいぅうぇ...
-1	108	62	1, 2, 1, 2, 3, 4, 5, 6, 7, 8, ...	ぁあぁあぃいぅうぇえぉおかがきぎくぐけげ...
-2	99	63	2, 3, 4, 5, 4, 5, 6, 7, 8, 9, ...	あぃいぅいぅうぇえぉおかがきぎくぐけげこ...
-20	91	94	1, 2, 3, 4, 5, 6, 7, 8, 9, 10...	ぁあぃいぅうぇえぉおかがきぎくぐけげこご...
-40	41	124	1, 2, 1, 2, 3, 4, 7, 5, 6, 9, ...	ぁあぁあぃいぇぅうぉえがかおけぐくぎげき...
-42	3	144	1, 2, 3	ぁあぃ

計算、合計し、これを全ての学生に対して行った。レーベンシュタイン距離の計算コストが低いため、重い計算ではなかった。

このようにして計算したレーベンシュタイン距離を元に学生を並び替えた、授業中の学生の電子ブック操作イベントの時系列データを図 1 に示す。横軸は授業時間の 0–90 分で、水平に同一学生の時系列イベントデータがその時のページ番号に従った色分けで表示されている。縦軸は、表 2 に示すように、シーケンス長とレーベンシュタイン距離により学生を並び替えている。先に示した方法で選んだ基準となる学生を y 軸 0 に配置しており、この授業中のレーベンシュタイン距離は全てこの学生のページ遷移シーケンスに対して計算する。y 軸正の方向には、シーケンス長が基準となる学生より長い場合のレーベンシュタイン距離の昇順に、y 軸負の方向には、シーケンス長が基準となる学

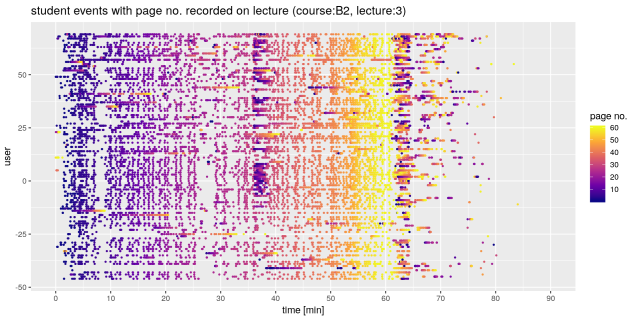


図 2 図 1 と同一教員、同一内容の別のコースでの授業 (course:B2, lecture:3)。

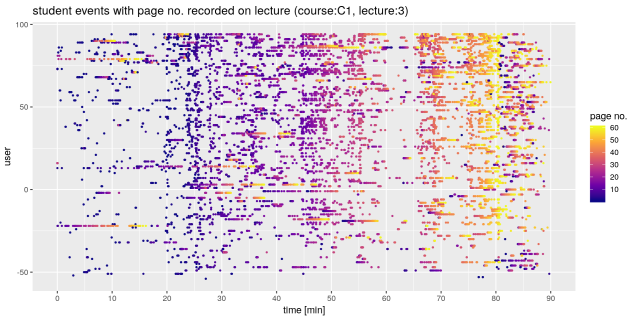


図 3 図 1 と同一教材を用いた別の教員の授業 (course:C1, lecture:3)。

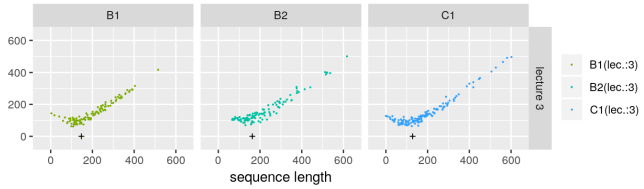


図 4 図 1–3 のシーケンス長とレーベンシュタイン距離の関係。

生より短い場合のレーベンシュタイン距離の降順に並べている。すなわちレーベンシュタイン距離は基準学生を中心に離れれば離れるほど大きくなり、シーケンス長は y 軸正方向におよそ増加することになる。

表 2 のシーケンス文字列は同表のページシーケンスに基づき、UTF-8 のコード表に割り当てて文字列化したもので、UTF-8 に対応した一般的レーベンシュタイン距離のライブラリが利用できるようにするためである。即ち、y 軸零のラインから離れるにつれて、シーケンスパターンの違いが大きくなっているはずである。

図 2 に、同じ教員 B が実施した他のコースでの状況を示す。弱冠の違いはあるが、似たパターンになっている。

図 3 に、別の教員 C が実施した同一内容のコースの様子を示す。図 1 図や 2 とはかなり異なるパターンを示している。

図 4 に、図 1–3 に対応するシーケンス長とレーベンシュタイン距離の関係を示す。表 2 からもわかるように、基準学生のシーケンス長の前後でレーベンシュタイン距離がおよそ逆転する。例えば、シーケンス長零の場合を考えると

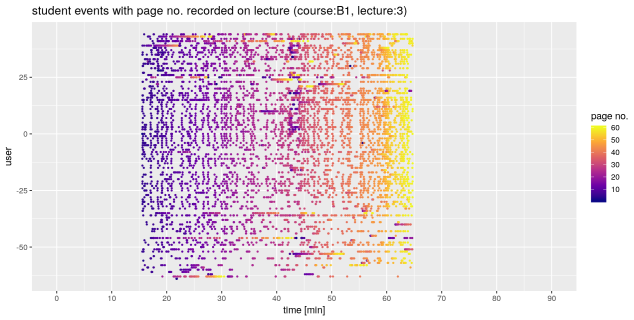


図 5 図 1 について担当教員のシーケンスとのレーベンシュタイン距離を用いた場合 (course:B1, lecture:3)。

1 文字ずつの変更操作が基準学生の文字列分必要になること、シーケンス長が非常に長い場合は少なくとも文字数の差分の変更操作が必要であることから理解できるであろう。

3.1.2 担当教員のデータを使用する場合

担当教員のシーケンスデータを学生のレーベンシュタイン距離を計算するのに使用する場合は、単純に各々の学生に関して担当教員のシーケンスデータとの間でレーベンシュタイン距離を計算すればよく、計算コストは殆どかからない。

ただし、この場合は注意すべきことがある。例えば、図 5 と表 3 に、図 1 と表 2 の場合のレーベンシュタイン距離の計算方法をこちらに変更したものを示した。図 5 の最初の 12 分位と最後の 25 分位のデータがないことがわかるであろう。実は、この時間には教員はページ操作を行っていない。最初の部分は、スライドなしに説明されたのであろうか、多くの学生も操作していないが一部の学生は操作している。また、最後の部分は、小テストの実施時間で、学生が LMS 上で小テストを受けつつ適宜電子ブックのページを参照しているのであろうか、バラバラのページを閲覧している。図 2 や図 3 にも同様の領域があることがわかる。

今回は、その授業が始まって、教員が最初のページを表示してから、最後のページを最初に表示した時点までを有効範囲として選択した。

表 3 を表 2 と比較すると、y 軸零の位置が大きく異なり、また、レーベンシュタイン距離の増え方も異なることがわかる。スライドを使った説明が始まる前の操作や、小テスト受験時の操作が大きく関わっているのではないかと思われる。

図 6 に、この場合のシーケンス長とレーベンシュタイン距離の関係を示すが、図 4 とおよそ同様の傾向である。ここで、C1 コースがないのは、教員データがないため、こちらの手法は使えないためである。

3.2 シーケンス長と小テストの得点

図 7 に、担当教員のシーケンスデータを使わない場合の、シーケンス長とその授業での小テストの結果の関係を示す。小テストは 5 点満点で、よい成績の学生が多い。こ

表 3　図 5 に関するデータ (抜粋)。

user (縦軸)	シーケンス長	レーベンシュタイン距離	ページシーケンス	シーケンス文字列
44	462	398	5, 6, 7, 8, 9, 8, 9, 10, 11, ...	ぅうぇえぉえぉおかがきぎくぎくぐけぐけげ...
40	224	168	6, 5, 6, 7, 8, 9, 10, 11, 12...	ううぅええぉおかがきぎくぐくぐけぐけげこ...
20	153	97	6, 7, 8, 9, 8, 6, 7, 5, 6, 7, ...	うぇえぉえうぇぅうぇえぉおかがきぎくぐけ...
2	110	58	7, 8, 6, 7, 5, 4, 3, 4, 5, 6, ...	ぇえうぇぅいいいぅうぇえぇえぉおかおかが...
1	111	47	6, 5, 4, 5, 6, 7, 8, 9, 10, 9, ...	うういぅうぇえぉおぉおかがおかがきぎくぐ...
0(教員)	109	0	1, 2, 3, 4, 5, 6, 5, 7, 8, 9, ...	ぁあぃいぅうぅえぇえぉおぉおぉおかおかがき...
-1	81	47	5, 6, 7, 6, 7, 6, 7, 8, 9, 10, ...	ぅうぇうぇうぇえぉおかおかがきぎくぐけげ...
-2	76	49	5, 6, 5, 6, 7, 8, 9, 10, 11, ...	ぅうぅうぇえぉおかがきぎくぐけげこござざ...
-20	76	59	5, 6, 7, 8, 9, 8, 9, 10, 11, ...	ぅうぇえぉえぉおかがきぎくぐけげこござざ...
-40	93	75	5, 6, 7, 8, 7, 8, 9, 10, 11, ...	ぅうぇえぇえぇおかがかがきぎきぎきくぐけ...
-60	48	88	5, 6, 7, 8, 9, 10, 11, 12, ...	ぅうぇえぉおかがきぁぃあいえぅぇぉうおか...
-64	3	106	1, 2, 3	ぁあぃ

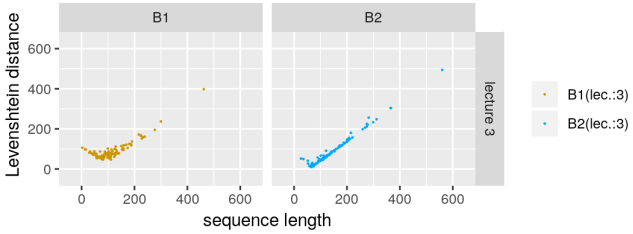


図 6　図 5 等のシーケンス長とレーベンシュタイン距離の関係。

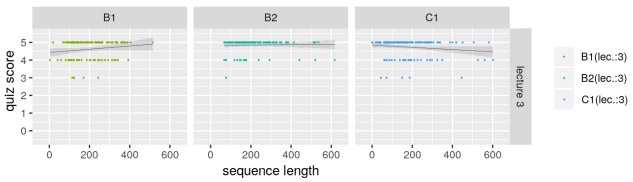


図 7　図 1–3 のシーケンス長と小テストの関係。

こでは紙面の関係から示せないが、他のデータも含めて、あまり相関はなさそうであった。これ以降、直線と振れ幅を示す場合、R の lm() 関数を用いた直線回帰で、振れ幅は 95%信頼区間を示す。

また、担当教員のシーケンスデータを用いた計算でも、あまり相関はなさそうであった。

図 8 は、8 回の授業の平均シーケンス長と小テスト合計点の関係を示したもので、平均シーケンス長が零付近で合計得点下がる傾向が見れる。これは、いつもページ遷移操作をしていない学生、例えば休みの多い学生が、成績が悪いことを示しているのではないかと思われる。これ以降、曲線と振れ幅を示す場合、R のデフォルトの局所多項式回帰 loess を用い、振れ幅は 95%信頼区間を示す。

担当教員のシーケンスデータを用いた計算でも同様の結果であった。

3.3　レーベンシュタイン距離と小テストの得点

図 9 に示すように、各回の授業に関してレーベンシュタイン距離と小テストの得点の関係を調べたが、シーケンスデータと成績の関係同様、相関があるようには見えなかった。なお、教員のシーケンスデータを用いた場合も同様であった。

図 10 に、8 回の授業の平均レーベンシュタイン距離と小テスト合計点の関係を示すが、シーケンス長の場合と違って、相関がなさそうであった。

そこで、レーベンシュタイン距離を、図 4 や図 6 にあるように、極小を挟んで、それ以下とそれ以上に分けて考えることとした。即ち、レーベンシュタイン距離の対象者のシーケンス長を基準値とし、それ以上と以下に場合分けする。図 11 に教員のページ遷移シーケンスを使わない場合を、図 12 に使う場合を各々示す。両図とも (a) に基準以下、(b) に基準以上の場合を示す。

(a) 基準以下の場合、レーベンシュタイン距離が大きくなるということは、およそシーケンス長が短くなることが、図 4 や図 6 からわかる。よって、図 8 と同様に、シーケンス長が短く零に近いところではテストの合計点が下がるのではないかと思われる。ただし、基準値付近で少し上昇しているように見える。こちらは、教員ないし基準学生より少しページ遷移が少ないほうが成績が上昇することを表しており説明が難しい。今後原因究明を行いたい。

(b) 基準以上の場合、レーベンシュタイン距離が大きくなるということは、教員ないし基準学生と違ったページを多く見ている学生で、自ら色んなページを見ていることを意味し、成績が上昇しているのではないかと思われる。

4.　まとめ

電子教科書を用いた面接授業における学習者のページ遷移に対して、レーベンシュタイン距離を用いた分析を試みた。授業中の電子テキストの操作イベントの記録データから、ページ滞在時間を無視したページ遷移のシーケンスと、そこからレーベンシュタイン距離を求めた。学生のレーベンシュタイン距離を求める対象として、担当教員のシーケンスを用いる場合と、その授業における全員のレーベンシュタイン距離の合計を最も低くする学生を選び、そのシーケンスを用いる場合について調べた。後者は、教員のシーケンスデータを必要とせず、自習や小テスト受験等、教員のシーケンスデータがない場合もレーベンシュタイン距離を求めることができる。

求めたレーベンシュタイン距離に関して、まだ予備的な

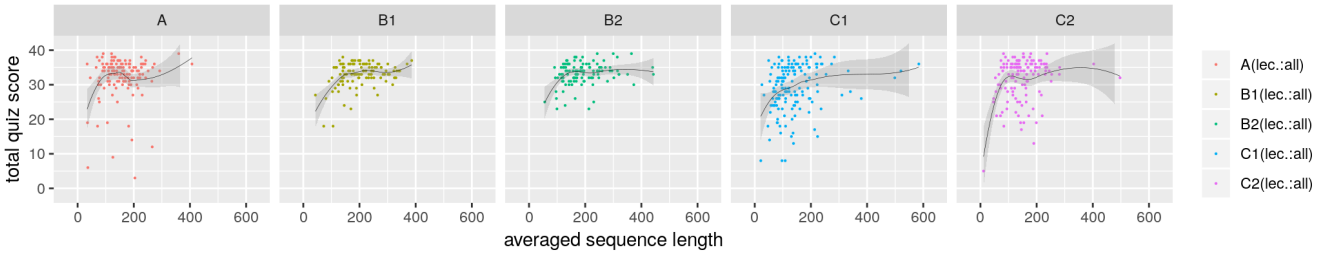


図 8 8 回の授業の平均シーケンス長と小テスト合計点の関係。

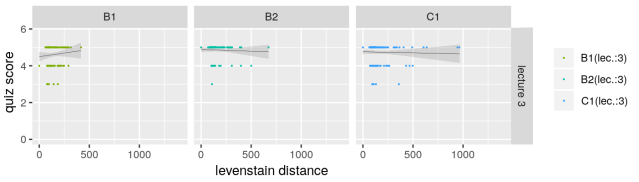


図 9 図 1–3 のレーベンシュタイン距離と小テストの関係。

分析の段階ではあるが、個々の授業回の中ではレーベンシュタイン距離と成績の間の関係は見いだせていないが、集計データに関しては、レーベンシュタイン距離を基準値 (対象となるシーケンスの長さ) の上下に分けることで、ある程度の関連性が見えた。基準値以下では、レーベンシュタイン距離が大きくなるにつれて一旦小テストの合計点が上昇しその後、下降する。基準値以上では、レーベンシュタイン距離が大きくなるにつれて合計点が上昇する。

現状では各回の授業データのみではダメで集計データを用いなければならない点、まだ十分に説明できない点がある点等、今後も検討を続けたい。また、小テストの得点分布、教員間の差異、レーベンシュタイン距離の時間変化、レーベンシュタイン距離以外のシーケンス比較等も今後検討したい。

謝辞

本研究では、日本教育工学会 SIG「教育・学習支援システムの開発・実践」第 10 回研究会で九州大学から提供された教育データを使用させていただきました [5,6]。ここに感謝の意を表します。また、本研究は JSPS 科研費 15H02795 の助成を受けたものです。

参考文献

[1] N Sclater, A Peasgood, J Mullan : Learning analytics in higher education, Jisc. Accessed February, 2016.
[2] 緒方広明, 殷成久, 毛利考佑, 大井京, 島田敬士, 大久保文哉, 山田政寛, 小島健太郎 : 教育ビッグデータの利活用に向けた学習ログの蓄積と分析, 教育システム情報学会誌, vol. 33, no. 2, pp. 58-66 (2016).
[3] 近藤伸彦, 畠中利治 : 学士課程における大規模データに基づく学修状態のモデル化, 教育システム情報学会誌 vol.33, no.2, pp.94–103 (2016).
[4] Vladimir Iosifovich Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals, Soviet Physics Doklady, 10(8), pp.707–710, 1966.
[5] Hiroaki Ogata, Chengjiu Yin, Misato Oi, Fumiya Okubo, Atsushi Shimada, Kentaro Kojima, Masanori Yamada. E-Book - based Learning Analytics in University Education, The 23rd International Conference on Computers in Education, pp.401-406, 2015.
[6] Hiroaki Ogata, Yuta Taniguchi, Daiki Suehiro, Atsushi Shimada, Misato Oi, Fumiya Okubo, Masanori Yamada, Kentaro Kojima. M2B System: A Digital Learning Platform for Traditional Classrooms in University, 7th International Conference on Learning Analytics & Knowledge, pp.155-162, 2017.
[7] Marcelo Worsley and Paulo Blikstein, Leveraging multimodal learning analytics to differentiate student learning strategies, In Proceedings of the Fifth International Conference on Learning Analytics And Knowledge (LAK '15), pp.360–367, 2015.
[8] Mina Shirvani Boroujeni and Pierre Dillenbourg, Discovery and temporal analysis of latent study patterns in MOOC interaction sequences, In Proceedings of the 8th International Conference on Learning Analytics and Knowledge (LAK '18), pp.206–215, 2018.
[9] RStudio Home: https://www.rstudio.com/products/RStudio/ (2018 年 11 月確認)

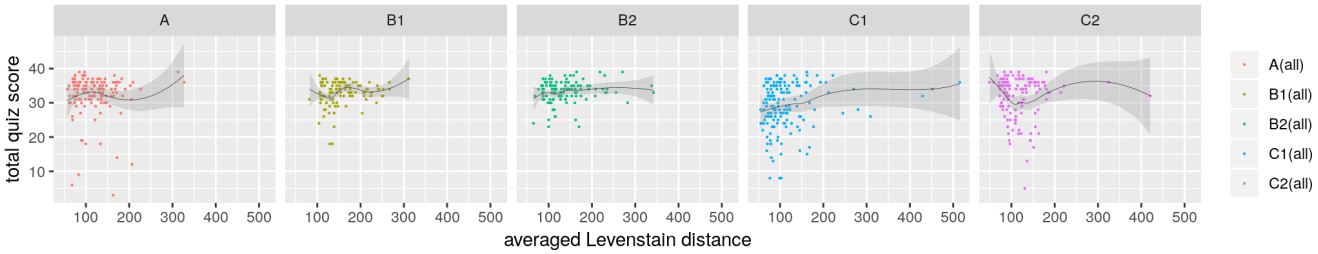
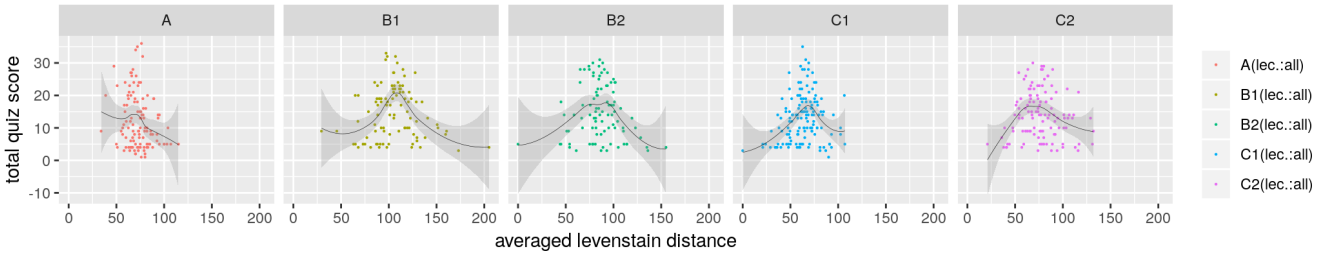
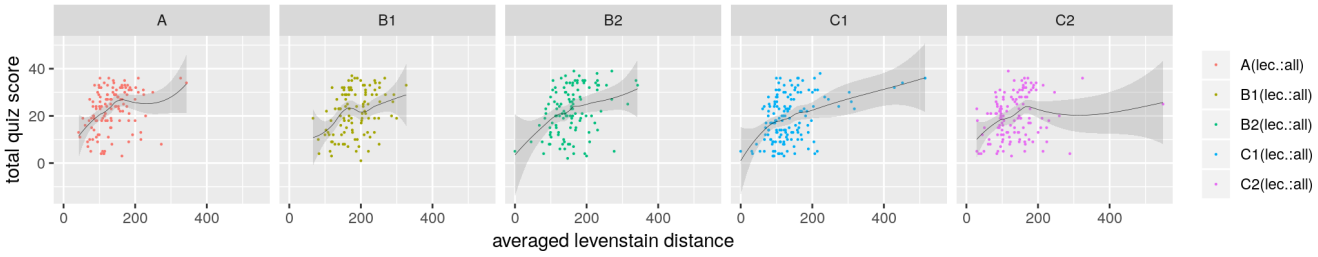


図 10 8 回の授業の平均レーベンシュタイン距離と小テスト合計点の関係。

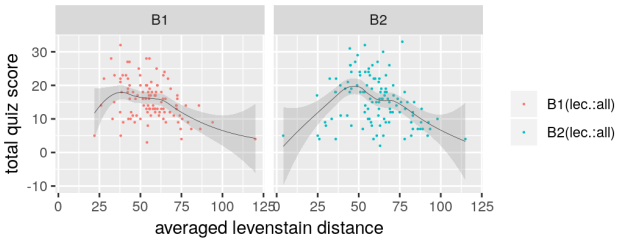


(a) シーケンス長が基準値以下の場合

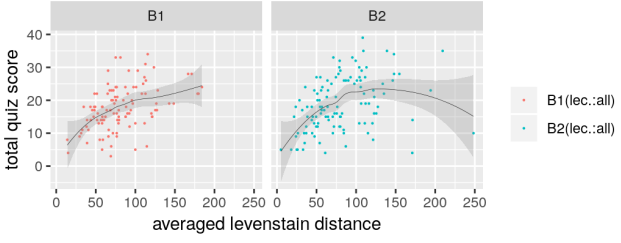


(b) シーケンス長が基準値以上の場合

図 11 8 回の授業の平均レーベンシュタイン距離と小テスト合計点の関係。教員のシーケンスを使わない場合。



(a) シーケンス長が基準値以下の場合



(b) シーケンス長が基準値以上の場合

図 12 8 回の授業の平均レーベンシュタイン距離と小テスト合計点の関係。教員のシーケンスを使う場合。