CLOSENESS ANALYSIS OF STUDENTS’ WORD-USAGE LOCATED IN A HIGH-DIMENSIONAL SPACE

Toshiro Minami and Yoko Ohura

Kyushu Institute of Information Sciences
6-3-1 Saifu, Dazaifu, Fukuoka 818-0117 Japan

ABSTRACT

As a part of our studies in investigating the relations between university students’ performances and their attitudes to learning, we propose a new method of representing word-usage profiles of students by embedding the word-usage vectors into a high-dimensional Euclidean space in this paper. The original source of data are the answer texts of students for a term-end evaluation questionnaire in a university class. By analyzing the relations in a couple of layouts of the word-vectors that are supposed to represent a kind of attitudes of students, we intend to find some kinds of relations between word-usages and academic achievements of students. The major achievements of this paper include (1) to investigate the relations of students’ word-usages themselves as well as (2) to develop new methods that clarify the differences of students in terms of their viewpoint in comparison with their achievements through these layouts.

KEYWORDS

Educational Data Mining (EDM), Natural Language Processing (NLP), Word-Usage Analysis, Graph Drawing, Simplicial Complex, High-Dimensional Euclidean Space

1. INTRODUCTION

Our major concern in our studies in the field of educational data mining (EDM) (Ames et al., 1988)(Ames, 1992)(Romero et al., 2007)(Silva et al., 2017) is to have more appropriate image about the students who attend our classes in universities. Based on the image for students’ learning attitudes, we are able to teach/guide/advice them more appropriately and motivate them for better learning. According to our observation as professor in some universities in Japan, quite a lot of students are not sufficiently motivated in learning and their attitudes to learning need to be improved in order to have more effective achievements as outcome.

It is often pointed out that many university professors lack teaching skills, probably including eagerness in education. We admit that this fact is true in one side. At the same time, it is true that students need to be more motivated in learning in another side. As a part of educational skill by professors, we are interested in understanding why such a number of students lack motivation toward learning. In order to understand more about students in their motivation, attitude, curiosity, interest range, etc. for learning, we have been conducting a series of study in analyzing data that can be obtained in our own experiences in teaching at classrooms; E.g. (Minami et al., 2017)(Minami et al., 2020)(Minami et al., 2022).

Our approach is different from many studies in the field of EDM, where quite a lot of studies use the target data which are obtained systematically or automatically from, for example, learning management systems (LMSs) so that the amount of data is big enough to be able to analyze them with the methods for big data (BD). For example, (Romero et al., 2008) gave a comparative study of data mining algorithms for classifying students using data from e-learning system and its major concern is to predict the student's outcome. In our approach, on the other hand, we aim to deal with student’s psychological tendency in learning, such as eagerness, diligence, seriousness. Another feature is that we intend to deal with small data (SD) obtained in our everyday classes rather than to deal with a huge amount of big data (BD) (Faraway et al., 2018)(Kitchin et al., 2015).

As a part of studies in our approach to small data analysis, we have been investigating retrospective evaluation texts of students by looking-back the course at a term-end class; E.g., (Minami et al., 2014)(Minami et al., 2015)(Minami et al., 2015a)(Minami et al., 2017)(Minami et al., 2022). As a result, we found that the students who have broader viewpoints in learning tend to have better achievements than those who have
narrower viewpoints in learning. In other words, the students who are able to locate the new knowledge in their knowledge network what they have already learned are tend to be able to get better achievements, or the term-end examination for evaluating what they have learned in the course.

We have also studied in analyzing class data with the seat occupation data in the classroom and we found that a student’s seat positions in a classroom relate to their outcomes (Minami et al., 2018). In this paper, we deal with text data of the same course as was used in the papers (Ohura et al., 2018) (Minami et al., 2017). By combining results obtained in analysis of both types of data, we can understand more about students on their viewpoints and their styles of study.

The rest of this paper is organized as follows: In Section 2, we describe the outline of the target data to be analyzed in this paper. In Section 3, we conduct the data analysis. First, we show how the data are processed from the original text data that are obtained as answer to a looking-back questionnaire to the word vectors, where each vector is considered to represent the corresponding student in her word-usage profile. Then, the distances between word-usage profile data are calculated and used for layout the relationship of students. We discuss the relations between the location of students in layout and their achievement in learning. Finally in Section 4, we conclude our discussions in this paper and prospect our future plans.

2. DATA

The data used in this paper came from a course in 2009 named “Exercise for Information Retrieval” in a junior (2-year) college (Minami et al., 2022). The students who were attending the class are in their second year and are going to graduate. The number of registered students was 35. The course was compulsory in order to get certificate for librarian. Thus, the students of this course were generally more motivated than other courses.

The major aim of the course was to let students become expert information searchers so that they had knowledge on information retrieval, and had skills in finding appropriate search engine sites and search keywords by understanding the aim and the background of the retrieval. The course consists of 15 lectures.

The term-end examination of the course consisted of 3 questions. The aim of these questions was to evaluate the skills on information retrieval, including the skills for planning and summarizing. These skills are supposed to have learned and trained in the course, through their exercises in the classes and while they do homework. We consider the score of term-end examination as a measure for student’s academic performance/achievement.

We also asked the students to evaluate themselves. The evaluation consisted of 12 questions; half of which are for evaluating the course, such as (Q1) What do you have learned in the course? Do you think they are useful? (Q2) What are the good points of the lectures? (Q3) What are the bad points that should be corrected? (Q4) Score the course as a whole, with the numbers from 0 to 100. The rests are evaluating students themselves: (Q6) What are the good points of the student herself regarding learning attitudes and efforts in learning during the course period? (Q7) What are the bad points that should be corrected of the student herself? (Q11) Score the student herself by considering her efforts and attitude in the course with the numbers from 0 to 100.

The amount of data used in this paper is very small. Therefore, it is impossible to extract useful information which is applicable in other classes. Our main aim of analysis of these small data is to find not only candidate findings that may be able to extend to other cases, but also to develop new analysis methods as many as possible as the first step to data analysis of lecture data we use. Then, we would apply the methods found in the first step to the data obtained from other classes. The methods are supposed to be evaluated according to their applicability to other classes, usefulness of the extracted information for lecturers in advising students.

3. ANALYSIS

In this section, we conduct analysis of the data which are described in Section 2. What we intend to do is to layout the closeness between a student and another student, and between a student and a group of students according to their word-usage profiles.
3.1 Outline of the Data Processing

Figure 1 shows the outline of the data processing in this paper. The process starts with the texts obtained as the answer to the questionnaire described in Section 2. They are converted to the corresponding word list using MeCab (MeCab). MeCab is a well-known morphological analyzer for Japanese texts. As the words in a sentence are not separated by spaces unlike in English, we need to use a morphological analyzer in order to deal with a word as a unit for processing. We are able to obtain the list of words with their part-of-speech (POS) information as the output by MeCab.

Then we count the number of occurrences, or frequency, of each word and create a word vector from the word list. A word vector consists of the corresponding frequencies of the words. Lastly, we calculate the mutual distances, and similarities, between word vectors. We consider the word vector the representation of the word-usage profile of the corresponding student in this paper. Thus, the distance between two word-vectors is considered to be the distance of the two students in terms of their word-usage profiles. Cosine similarities between vectors are popularly used in order to measure the amount of closeness, which is not only easy to recognize intuitively, but also easy to calculate; by using inner product of vectors, i.e., component-wise product and summing them up.

However, we do not use cosine similarity directly for the index of closeness between students. As we can easily estimate that the usage of words would not be very different among students; frequently used words might be used frequently by all students and small number of words might be used differently by students to students. Thus, we use sine distance because its values are more different than cosine similarities for small angles. Also, sine distance can be calculated by using cosine similarities and easy to calculate as well.

In this section, we use two methods for drawing which show the layout how the students are close each other. The methods are (1) drawing layout by using graph drawing algorithm and (2) drawing layout by projecting into 2-dimensional plane after arranging the targets into a high-dimensional Euclidean space.

3.2 Calculating Distances between Students/Groups

We describe how to define the groups of students and how to represent word-usage. We also show how the distances are calculated in detail in this subsection. Let S be the set of students and W be the set of words. Let ans(s) be the text which is obtained as the answer to the questionnaire by the student s. Then ans(s) is of the form w1,w2,…,wn, where the number of the words k=|ans(s)| is determined by the student s.

We get the word-list w*=[(w1,pos(w1)), (w2,pos(w2)),…,(wk,pos(wk))] as the output of MeCab by giving ans(s) as input, where pos(w) is the part-of-speech (POS) of the word w. In this paper, however, we omit the pos part in the word list. Thus, we consider w*=[w1,w2,…,wn]. We denote the i-th element of the list w* by w*[i]; thus, w*[i]=wi.

Next, we count the frequency of words in a word list. The word vector w# of w* is defined as follows:

\[ w# = [\text{freq}(w_1, w*), \text{freq}(w_2, w*), \text{freq}(w_3, w*), \ldots, \text{freq}(w_k, w*)] \]

where n is the number of words in W (n=|W|) and W=[w1, w2, …, wn], and freq(w,w*) is the number of occurrences of the word w in the word list w*; thus, freq(w,w*)=[|w*[i]|=n, 1≤i≤k]).

Now, we define the sine distance between two students using the corresponding word-vectors. Let w#1 and w#2 be the word vectors of student s1 and s2, respectively. Firstly, we adjust the lengths of the vectors to 1. Since the number of words used by a student varies a lot from student to student. Some students answer with a lot of words, whereas some students answer very shortly. In order to compare the word-usage profiles of...
students, we have to adjust the length of the word vectors so that the component values represent some kind of ratios of words used by the student. Now, we define $w_i'$ for $i=1, 2$ by:

$$w_i' = \frac{w_i}{\|w_i\|},$$

where $\|w_i\|$ is the length, or norm, of $w_i$, where norm is defined by $\|w\| = \sqrt{\sum_{j=1}^{w_i} [w_i[j]^2]}$, i.e., the length in the Euclidean space. Then, we define the cosine similarity of two vectors $w_1'$ and $w_2'$ by: $\text{CosSim}(w_1', w_2') = w_1' \cdot w_2'$, where $\cdot$ denotes the inner product $\sum_{j=1}^{w_i} w_1'[j]w_2'[j]$. Sine distance is defined by: $\text{SinDis}(w_1', w_2') = \sqrt{1 - \text{CosSim}(w_1', w_2')^2}$ and distance of students by: $d(s1, s2) = \text{SinDis}(w_1', w_2')$.

Figure 2. Layout of Students using graph drawing method by spring model

Figure 2 shows the layout of students using distances in this definition between students drawn by a graph drawing algorithm. A circle represents a student and the upper label shows the student’s name, or ID, and the lower label shows the achievement score, i.e., the examination score, of the student. The color of the circle shows the group of the student based on the achievement score so that it becomes easy to see which group she belongs to. The edges that connect two students show that the distances between the students that the edge connects has higher closeness, i.e., small distance, than other pairs of students. As we observe Figure 2, we cannot say that students having similar achievement scores are close each other in their word-usages in general. On the other hand, some groups seem to be more separated than other groups in the graph. We would like to investigate more about the relationship between achievement scores and word-usages in the following sections.

3.3 Group-based Analysis using Graph Drawing Algorithm

As we have observed that the achievement scores and word-usages have little correlations in student to student, we divide the students into groups according to the achievements and see if the groups have some kinds of difference in word-usages. Firstly, we show the outline of grouping of students in this subsection. Table 1 shows outline of grouping. Students are divided into 5 groups, namely G1 to G5, according to their achievement scores so that each group contains 7 students. The achievement scores range from 79 to 99 and the mean score is 84.1 for G1, for example.
Table 1. Outline of Grouping

<table>
<thead>
<tr>
<th>Group</th>
<th>Coloring</th>
<th>Number of Students</th>
<th>Minimum Score</th>
<th>Mean Score</th>
<th>Maximum Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>Red</td>
<td>7</td>
<td>79</td>
<td>84.1</td>
<td>99</td>
</tr>
<tr>
<td>G2</td>
<td>Purple</td>
<td>7</td>
<td>75</td>
<td>75.9</td>
<td>78</td>
</tr>
<tr>
<td>G3</td>
<td>Green</td>
<td>7</td>
<td>64</td>
<td>68.6</td>
<td>74</td>
</tr>
<tr>
<td>G4</td>
<td>Light Blue</td>
<td>7</td>
<td>54</td>
<td>58.1</td>
<td>63</td>
</tr>
<tr>
<td>G5</td>
<td>Black</td>
<td>7</td>
<td>27</td>
<td>40.6</td>
<td>53</td>
</tr>
</tbody>
</table>

The word vectors of a group are calculated by just adding up all the number of word occurrences of the member of the group and the distances from a group to another group and a student are calculated in the same way as in calculating between students as has been described.

Figure 3 shows the layout of groups using the graph drawing method by spring model as was used in Figure 2. The edge labels show the distances between the groups. As we can see in Figure 3, the maximum distance value is 0.33 between G4 and G5, whereas the minimum distance value is 0.21 between G1 and G3. The mean value of a group to other 4 groups of G1 to G5 are 0.25, 0.27, 0.26, 0.31, and 0.30, respectively. Thus, roughly speaking, G1 is the group which is the mostly close to other groups, and the distances increase in the order of G3, G2, G5, and G4. Possibly, G1 is located in the center of the groups because of this property. Another interesting fact is that G3, the middle group in terms of achievement score, is the one closest to G1.

Next, we would like to arrange students by using the distances of a student to these 5 groups. We also use the spring algorithm by fixing the layout of the groups as in Figure 3. Figure 4 shows the results. We have to point out that the length parameter for the strength of spring is given by using the cosine similarity value because the length should be larger as the distance becomes smaller. Further, we multiply the similarity value by 8 for emphasis so that the groups and the students are located not too-close each other and well-separated.

As we can see, the students in G1 are located somewhere near G1 except st33. For G2, the students are located rather away from G2 in the central area of the graph, near G1. Among them, st24 is far away from G2 and is located in-between G4 and G5. For G3, the students are located relatively wide areas in the graph. 2 of
them are close to G2, 3 of them are in-between G4 and G1, and the rest 3 are located relatively close to G3. For G4, the students are scattered in all areas of the graph. Among 7 students, 4 of them are located in the area in-between G4 and G5 and 3 others are located close to G1, G2 and G3 separated. For G5, students are also located separately in a wide area; st32 is located next to G5, st31 and st19 is in-between G5 and G2, st28 is very close to G2, st27 and st30 are close to G3, and st35 is located in-between G3 and G4.

3.4 Group-based Analysis by Embedding to Simplicial Complex in Euclidean Space

In this subsection, we propose a different approach to locate groups and students according to the distance data. As the graph drawing methods are supposed to arrange objects in 2-dimensional plane, 2 objects that have big distance may be located close to each other because of their distances to other objects. Thus, we cannot say that the two objects drawn close to each other are actually close to each other in terms of the distance value. Because of this, the observations from Figure 4, which are found in the previous subsection may be wrong, and thus we have to confirm them in different methods.

In the new approach, we realize the arrangement of objects (groups and students) so that their distances are preserved by embedding them into a high-dimensional Euclidean space as a simplicial complex (Tancer, 2011). As we cannot recognize how the objects are located in such a high-dimensional space, we project them into a 2-dimensional plane for visualization. Then, if we want to confirm the closely located two objects in the projected plane are actually close each other in the high-dimensional space, we can change the plane for projection so that we can see in a different angle and check if the two objects are still close to each other.

3.4.1 Simplicial Complex

Let X be the set of objects. We call K a simplicial complex of X if: (i) for any member x in K is a subset of X, and (ii) for any member x in K, any non-empty subset of x is also a member of K. We call a member of K a face of K and define the dimension of a face x by dim(x)=|x|-1. We call a face x as a vertex if dim(x)=0; i.e., a vertex is a face of K having only one element, i.e., x={v}. We identify {v} and v and deal with v as a vertex.

Our image for a simplicial complex with n-dimension is a kind of extended triangle. For example, the face {v1, v2} is a line segment that connects the vertices v1 and v2 which has dimension 1, and the face {v1, v2, v3} is a triangle having v1, v2, and v3 as its vertices and {v1, v2}, {v2, v3}, and {v3, v1} are its sides. The triangle’s dimension is 2, which is easy to recognize because the triangles are embedded in a plane, which is 2-dimensional. We call the simplicial complex which consists of all subsets of objects complete simplicial complex for O in the following sections.

3.4.2 Embedding a Simplicial Complex in a High-dimensional Euclidean Space

Let R be the set of real numbers and let d: O×O→R be the distance function which satisfies the following conditions: For any o1, o2, and o3 in O, (i) d(o1, o2)≥0, (ii) d(o1, o2)=0 if and only if o1=o2, (iii) d(o1, o2)=d(o2, o1), and (iv) d(o1, o2)+d(o2, o3)≥d(o1, o3). Note that the sine distance for the groups G={G1, G2, ..., G5} satisfies these conditions. We denote d for the sine distance for G in the following discussions.

3.4.3 Layout of Groups and Students in a Euclidean Space and their Projected Image

We can embed the simplicial complex K(G) in R^4 by assigning coordinates to each group in G. We denote cor(Gi) for the coordinate assigned to Gi in R^4, and d_{ij} for d(Gi, Gj) for the groups Gi and Gj in G. Then, Cor(Gi) is determined step by step as:

(i) cor(G1)=(0, 0, 0, 0)  
(ii) cor(G2)=(d_{12}, 0, 0, 0)  
(iii) cor(G3)=(x, y, 0, 0), where x and y are determined by solving the equations  
a. d(cor(G1), cor(G3))=d_{13}, or x^2+y^2=d_{13}^2 and  
b. d(cor(G2), cor(G3))=d_{23}, or (x+12)^2+y^2=d_{23}^2  
(iv) cor(G4)=(x, y, z, 0) such that x, y, and z satisfy the equations d(cor(Gi), cor(G4))=d_{i4} for i=1, 2, and 3.
\( \text{cor}(G_4) = (x, y, z, v) \) such that \( x, y, z, \) and \( v \) satisfy the equations \( d(\text{cor}(Gi), \text{cor}(G5)) = d_i \) for \( i = 1, 2, 3, 4 \).

As the result, we have the following coordinate assignment for the groups in \( G \). (i) \( \text{cor}(G1) = (0, 0, 0, 0) \), (ii) \( \text{cor}(G2) = (0.23, 0, 0, 0) \), (iii) \( \text{cor}(G3) = (0.09, 0.19, 0, 0) \), (iv) \( \text{cor}(G4) = (0.11, 0.04, 0.28, 0) \), and (v) \( \text{cor}(G5) = (0.06, 0.04, 0.06, 0.24) \).

We may identify a group name and its coordinate.

As the result, we have the following coordinate assignment for the groups in \( G \). (i) \( \text{cor}(G1) = (0, 0, 0, 0) \), (ii) \( \text{cor}(G2) = (0.23, 0, 0, 0) \), (iii) \( \text{cor}(G3) = (0.09, 0.19, 0, 0) \), (iv) \( \text{cor}(G4) = (0.11, 0.04, 0.28, 0) \), and (v) \( \text{cor}(G5) = (0.06, 0.04, 0.06, 0.24) \).

We use weighted average for locating the students in \( R^4 \). Let \( s \) be a student and \( d(s, Gi) \) be the distance between \( s \) and group \( Gi \) for \( i = 1, 2, \ldots, 5 \), which we denote \( d_i \) for brevity.

Let \( d_i' = \frac{d_i}{\sum_{j=1}^{5} d_j} \).

In order to calculate the weighted average location for \( s \), we have to define the weight \( w_i \) for \( Gi \) (\( i = 1 \) to 5) so that \( w_i \geq 0 \) and \( \sum w_i = 1 \). Then we can define \( \text{cor}(s) = \sum w_i \times Gi \).

Unfortunately, distance and weight relate inversely; weight increases as distance decreases, and vice versa. We define weight \( \frac{-\log(\text{distance})}{\text{emph}} \) for inversing their values by considering that distance has a positive value with maximum of 1. Thus, the weight ranges from 0 to infinity.

Further, we use the emphasized weight by powering emphasis factor \( (-\log(\text{distance}))^{\text{emph}} \) so that the students are well-distributed. In our case we decided to use \( \text{emph}=16 \) after we tried several values.

Lastly, we have to decide how to visualize the layout in \( R^4 \). We choose the projection to a 2-dimensional plane. The plane is the one spanned by a pair of given 2 vectors in \( R^4 \) so that it contains the origin \((0,0,0,0)\).

Let \( v_1 \) and \( v_2 \) be vectors in \( R^4 \). We make an orthogonal coordinate system by giving two unit-vectors \( u_1 \) and \( u_2 \) as follows: \( u_1 = v_1 / ||v_1|| \) and \( u_2 = u_2' / ||u_2'|| \) where \( u_2' = v_2 - (v_2 \cdot u_1) u_1 \) (Note that the dot operator \( \cdot \) is the inner product). Note that \( u_1 \) and \( u_2 \) are unit vectors by definition and they are orthogonal because their inner product is 0.

By using \( u_1 \) and \( u_2 \), the x- and y-coordinates on the projected plane of a vector \( v \) in \( R^4 \) is given by: \((x,y) = (v \cdot u_1, v \cdot u_2)\).

Figure 5 (left) shows the projected layout of the groups and students by giving \( v_1 = G_2 - G_1 \) and \( v_2 = G_3 - G_1 \). Note that \( G_1, G_2, \) and \( G_3 \) here are coordinates in \( R^4 \) and \( G_2 - G_1 \) and \( G_3 - G_1 \) are the vectors of \( \text{cor}(G2) - \text{cor}(G1) \) and \( \text{cor}(G3) - \text{cor}(G1) \), respectively. Figure 5 (right) shows the projected layout by giving \( v_1 = G_3 - G_1 \) and \( v_2 = G_5 - G_1 \).

As we can see that st11 locates very close to G2 in both figures, we may be able to recognize st11 and G2 are located very close to each other in \( R^4 \). Similarly, st18 locates very close to G3.

4. CONCLUDING REMARKS

In this paper, we have investigated the relation between word-usage of a student and her academic achievement. The source data are obtained as the answer text to a term-end questionnaire. Distances were calculated from
the word-usage vector. We considered the distance as the distance between a student and another student and the distance between a student and a group of students. The distances were drawn in two methods: (1) using graph-drawing algorithm and (2) projecting the layout in a high-dimensional Euclidean space into a 2-dimensional plane. Since the layouts of the students and groups of students are essentially high-dimensional, we cannot say the closeness of two objects just by looking in a visualized layout. We have to confirm by combining other method(s). In the second method we were able to change the plane for projection and see the relation in more precisely.

The topics to be investigated in the future include: (i) to explore further on the relation between the students’ achievements and their viewpoints that appear in their answer text, (ii) to apply the proposed method to other data, including the one we used for seat position analysis such as (Minami et al., 2018), (iii) to digitize more educational data we have in our lectures and what we can find in these data, and so on.

ACKNOWLEDGEMENT

This work was supported in part by JSPS KAKENHI Grant Number JP17K00502.

REFERENCES


Ames, C. and Archer, J.

REFERENCES

This work was supported in part by JSPS KAKENHI Grant Number JP17K00502.

ACKNOWLEDGEMENT

This work was supported in part by JSPS KAKENHI Grant Number JP17K00502.

REFERENCES


