短文レポート評価のための短文レポート群の特徴に基づく可視化システムの開発

山添 久稔　大囿 忠親　新谷 虎松

本研究では，講義後の理解確認のための短文レポートの評価を自動化するためのシステムの実現を目指している．本研究のチャレンジングな点は，正答および誤答をシステムを与えることなく，自動評価するための技術を実現することである．本稿では，自動評価のため，対象とする短文レポートの特徴より，正答と誤答を判定することを目的に，F1) レーベンシュタイン距離を元にした特徴量，F2) 頻出語を元にした特徴量，F3) 3-gram を元にした特徴量，という 3 つの特徴量を設計した．そして，3 つ特徴量に基づくレポート群可視化システムの開発を行った．また，3 つの特徴量の組み合わせの内，いずれの組み合わせが最も正答と誤答を判定できるかの実験を行った．実験の結果として，3 つ全ての特徴量を全て使った組み合わせが，最も適した組み合わせであった．

In this research, we aim to implement a system to automate the evaluation of short reports for checking student's knowledge. The challenging aspect of this research is to realize a technology for automatically evaluating correct or wrong answers by unsupervised learning. In this paper, we designed the following three features for automatic evaluation of correct answers or wrong answers by using the characteristics of the target short reports: F_1) A feature based on Levenshtein distance, F_2) A feature based on frequent word, and F_3) A feature based on 3-gram. We developed a report group visualization system based on these three a feature. We also conducted experiments to determine which combinations of the three feature are most appropriate for judging answers. As a result of the experiments, it turned out that the combination of all features is best.

1 はじめに

本研究では，講義後の理解確認のための短文レポートの評価を自動化するためのシステムの実現を目指している [3]．本研究のチャレンジングな点は，正答および誤答をシステムに与えることなく，自動評価するための技術を実現することである．ここでは，誤答となるレポートを正答からの外れ値とする，外れ値検出として考えることとした．

講義における短文レポートの評価は，評価者の大きな負担となる．そのため，短文レポートの評価を支援するシステムが必要である．大学の講義程度の短文レポートに対し，システムによる評価支援を試みる際，事前準備により評価者の負担を大きくしてはならない．そこで，本稿では，2 章で述べる特徴を持った短文レポート群に対象を絞り，この短文レポート群に対し，事前準備を必要としない短文レポート評価支援システムの開発を行う．

本稿でいう短文レポートの評価支援とは，1) 短文レポートの正誤を教師なしで判定，2) レポート群の特徴を 3 次元グラフとして可視化，および 3) 受講者へのフィードバックの 3 点である．本稿では主に 1) および 2) について説明する．

以降，本稿では．2 章で，対象とする短文レポートの特徴と関連研究との比較について述べる．3 章で，短文レポート評価支援システムの実装について述べ，4 章でシステムの構成図と実行例を述べる．また，5 章で評価実験を行い，6 章にて本稿をまとめる．

A Development of Visualized Evaluation System for Short Reports based on Caractaristics of the Reports

Hisatoshi Yamazoe, Tadachika Ozono, Tramatsu Shintani, 名古屋工業大学大学院情報工学専攻, Dept. of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology.

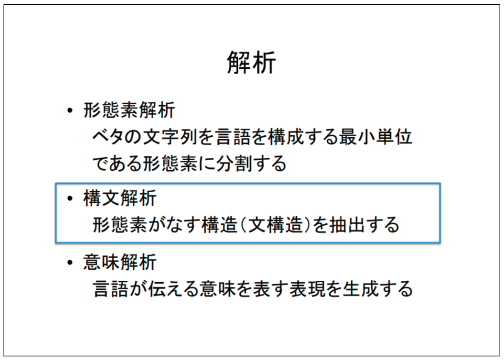

図 1　短文レポートの正答が書かれた講義資料

2　対象とする短文レポートの特徴と関連研究

2.1　対象とする短文レポートの特徴

本稿で対象とする短文レポートは，毎回の講義後に出題され，その目的は，講義受講者に講義内容の理解と復習を勧めるものである．講義受講者は，ウェブページより講義資料を自由に手に入れることができる．また，短文レポートの難易度は，講義資料の内容の一部がそのまま短文レポートの正答になっているような簡単なものである．そのため，本稿では，1) 解答間の類似度が高い，2) 正答が大多数を占める，3) 正答と誤答では含まれる語彙が異なる，という 3 つの特徴を，短文レポートの解答が持つと仮定する．

具体例として，本学の学部 3 年生向けの講義である「自然言語処理」の講義において，出題された短文レポートを 1 つ示す．短文レポートの問題文は,「自然言語処理とは？」というものである．また，図 1 は，短文レポートが出題された際の講義で用いられた講義資料の一部であり,「自然言語処理とは？」という問題に対する解答が示された講義資料である．講義資料では，自然言語処理を，形態素解析，構文解析，意味解析の 3 つに分け，それぞれについて概略の説明を行っている．また，実際の短文レポートの解答として,「自然言語処理とは，形態素解析，構文解析，意味解析の 3 つからなる．形態素解析とは，ベタの文字列を言語を構成する最小単位である形態素に分割するものである．構文解析とは，形態素がなす構造 (文構造) を抽出する．意味解析とは，言語が伝える意味を表す表現を生成する.」といった講義資料上の説明をそのまま繋げた解答が想定されている．

表 1　特徴量のラベル

ラベル	特徴量
F_1	レーベンシュタイン距離を元にした特徴量
F_2	頻出語を元にした特徴量
F_3	3-gram を元にした特徴量

本研究では，正答および誤答をシステムに与えずに正答判定を行うために，誤答となるレポートを正答からの外れ値とする．外れ値検出とかんがえることとした．ここで，上記のレポートに対し，tf-idf を用いてみたが，満足する結果を得ることができなかった．原因として，文章量が多いレポートの各形態素に対する tf の値が小さくなるためであると考えられる．そのため，各短文レポート間の文章量の差に依存しない特徴量の設計が必要不可欠である．

2.2　関連研究と本研究の比較

短文レポートの評価支援に関する研究は広く行われている．石岡らは，大学入試センター試験に代わる記述式の試験やその模擬試験などに対してアプローチしている [5]．石岡らの手法は評価の事前準備として，模範解答に加え，模範解答内の語彙についての言い換え表現，部分点，一部の回答を用いて作成した評価モデルの回答の評価を行う．また，短文レポートの評価結果を視覚的に表現する研究では，椿本らの研究が挙げられる [4]．椿本らの研究では，短文レポートの評価結果を円錐状の 2 次元上にマップし，短文レポート評価者への支援を行った．また，外れ値検出を文章に対して利用した研究として，Cha らの研究が挙げられる [1]．文献 [1] では，テキストマイニングに対し，Scholkopf らが開発した，One-Class-SVM [2] を用いた手法が提案されている．

3　正答／誤答を判定するための 3 つの特徴量と誤答判定機構

短文レポートに対する特徴量として，F_1) レーベンシュタイン距離を元にした特徴量，F_2) 頻出語を元にした特徴量，F_3) 3-gram を元にした特徴量の 3 つの特徴量を提案する．また，以下，それぞれの特徴量

を表 1 の通り表記する．この 3 つの特徴量は，それぞれ異なるヒューリティクスに基づき，1 回の短文レポート課題において，ある短文レポートがその他の短文レポートにどれほど近いかを特徴量として算出する．本システムが対象とする短文レポートが持つ特徴を考慮すると，算出された特徴量が 0 に近いほど，正解らしい答案であると期待することができる．この章では，3 つの特徴量と，3 つの特徴量を元にした誤答判定機構について述べる．

3.1 レーベンシュタイン距離を元にした特徴量

F_1 は，ある短文レポートと，他の短文レポートとのレーベンシュタイン距離を計算したものの和である．対象となる短文レポートは，例として示した「自然言語処理とは？」の解答のように，講義資料上の解答の部品を，自然言語として自然になるように，接続して解答を作成される．その際，正答解答集合には，句読点や「コンピュータ」,「コンピューター」等の一文字単位での表記揺れを除外すれば同一の文章となる解答が存在している．そのため，特徴を持つ短文レポート群を対象に，レーベンシュタイン距離の和を考える．類似した正答集合内の解答は，レーベンシュタイン距離の和が小さくなり，誤答集合内の解答はレーベンシュタイン距離の和が大きくなると期待できる．具体的な各短文レポートの F1) の算出方法を述べる．短文レポート集合 R の各短文レポート r_i のレーベンシュタイン距離を元にした特徴量の値 r_i^l は，式 (1) で表される．$lev(r_i, r_j)$ は，短文レポート r_i と短文レポート r_j のレーベンシュタイン距離である．

$$r_i^l = \sum_{r_j \neq r_i, r_j \in R} lev(r_i, r_j) \tag{1}$$

3.2 頻出語を用いた特徴量

F_1 は，文字単位の表記揺れをグラフ上の距離として表現した．しかし，レーベンシュタイン距離は，文章における文字の入れ替えにも影響する．具体的には，例で示した,「自然言語処理とは？」の短文レポートの場合，解答として，3 つの解析が列挙されたものを例示したが，同一の文で，それぞれの解析に対する説明文が入れ替わった際にも，レーベンシュタイン距離が大きくなり，誤答と文が入れ替わった正答との判別が困難となる．そのため，F_2 を提案する．F_2 は解答の語彙のみに着目した特徴量である．特徴を持つ短文レポートを対象とした場合，同様に正答の値が 0 に近づくことが期待できる．具体的な，F_2 に対する各短文レポートの値の算出方法を述べる．まず，短文レポート群の内，半数の短文レポートに出現する名詞を頻出語とし，この集合を名詞集合 W とする．そして，各短文レポート r_i に対し，次元 $|W|$ の，名詞の出現回数を要素とするベクトル v_i をつくる．最後に，ベクトル v_i の和の逆数を取る．この操作により，算出されたものが，頻出後を元にした特徴量である．

3.3 3-gram を元にした特徴量

3-gram 用いることで，名詞だけでなく，名詞に関わる助詞や副詞なども考慮にいれたレポート間の距離を考えることができる．3-gram を用いることで，F_2 では表現できない，名詞と副詞や助詞等のつながりを表現したレポート間の比較を行うことができる．具体的には,「を」,「に」等の格助詞と，格助詞の前の対象の 2 文字により，その文章において，対象の 2 文字が文章中の目的語なのかが表現できる．レポート r_i における，$F3$ の値 r_i^{tri} はレポート r_i より作られた 3-gram を $T(r_i)$ とすると式 (2) で求めることができる．

$$r_i^{tri} = \frac{1}{\sum_{r_j \neq r_i, r_j \in R} \left(\frac{|(T(r_i) \cap T(r_j))|}{\min\{|T(r_i)|, |T(r_j)|\}} \right)} \tag{2}$$

3.4 誤答判定機構

誤答判定機構とは，外れ値検知アルゴリズムを用いて，誤答らしい短文レポートを判定する機構である．誤答の判定は，以下の 4 ステップで行う．ステップ 1 では，評価対象の短文レポート群に対し，F_1,F_2,F_3 を求める．ステップ 2 では，求めた値を正規化する．ステップ 3 では，One-Class-SVM を用い，正規化を行った 3 次元データに対し，外れ値検知を行う．最後にステップ 4 では，外れ値と判定された短文レポートを誤答らしい短文レポートとして判定する. ただし，外れ値検知を行う際，短文レポートの文章量が，短文レポート群の文章量の平均より，1.5 倍以上大きいレ

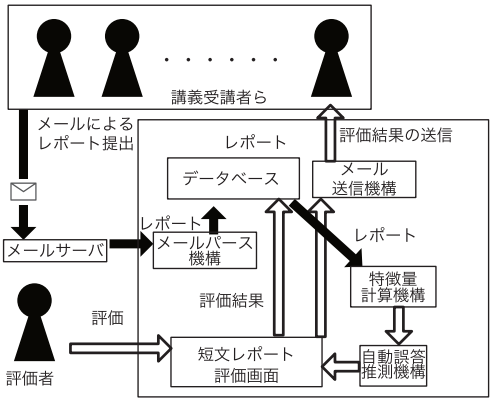

図 2　システム全体の構成図

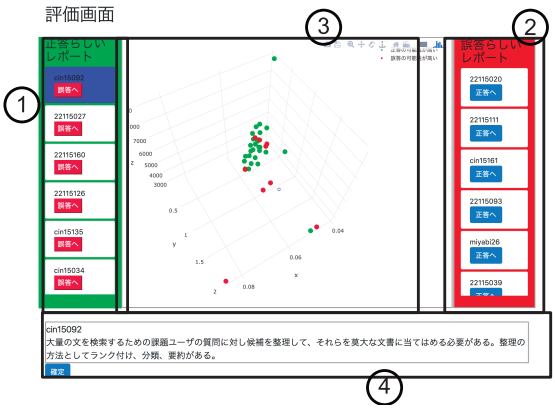

図 3　短文レポート評価画面

ポートに関しては予め正答と判定する．これは，長文の解答において誤答は極めて少ないためである．

正規化には，scikit-learn[†1] の preprocessing.scale/1 を，One-Class-SVM は，同様に scikit-learn の svm.OneClassSVM/3 をそれぞれ用いた．One-Class-SVM にはパラメータが nu，$kernel$，$gamma$ の 3 つ存在する．本システムでは，$nu = 0.3$，$kernel =$ “rbf”，$gumma =$ “$auto$” をそれぞれグリッドサーチにより設定した．

また，ステップ 2 において，正規化を行う際の 3 つの特徴量の組み合わせについて，F_i，$F_{i,j}$，$F_{1,2,3}$ という表記を行う．F_i は，F_i 単体で用いるという意味である．$F_{i,j}$ は F_i と F_j を用いて正規化を行うという意味である．$F_{1,2,3}$ は 3 つ全ての特徴量を用いるという意味である．これらの特徴量の組み合わせにおける，最適な組み合わせについては，5 章で比較実験を行う．

4　システムの構成図と実行例

システム全体の構成と，実際に短文レポートを評価する際の実行画面について述べる．

4.1　システム全体の構成図

システム全体の構成図を図 2 に示す．図 2 中の中抜き矢印は評価者が評価を行う際の流れを，塗りつぶし矢印はレポートの提出がなされた際の流れをそれぞれ示す．まず，講義受講者は指定されたメールアドレスに指定の様式で短文レポートの提出を行う．提出されたメールは，様式に従いパースされ，データベースに保存される．短文レポートの提出期限が過ぎた後，評価者は評価を行う．評価者は，提出期限後の任意のタイミングで，短文レポート評価画面を web ブラウザ上で開き，短文レポートの評価を行う．レポートの評価を行う際に，データベースより該当するレポート群が特徴量計算機構へと送られ，各レポートに対し特徴量が計算される．計算された値を用い，誤答判定機構により誤答が判定され，ユーザインターフェースである，短文レポート評価画面に送信される．評価結果は，データベースと講義受講者の元へ送信される．

短文レポートの提出方法を，専用の web ページではなく，メールにしたのには，以下の 2 つの理由がある．1 つ目の理由は，短文レポート提出に対して，ロバストなシステムにする必要があったためである．専用の web ページによる提出では，短文レポートの提出を行った際，サーバ側に何らかのトラブルが生じていた場合，提出した短文レポートが失われてしまう．しかし，メールによる提出では，万が一，サーバ側にトラブルが生じていた際でも，送信履歴として受講者側にもオリジナルの短文レポートが保存されているため，容易に再送することができ，受講者に再度 1 から短文レポートの記述を求めなくて済む．2 つ目の理由として，現在の対象としている短文レポートが

†1　http://scikit-learn.org/

表 2　誤答の検出

	使用した特徴量	F_1	F_2	F_3	$F_{1,2}$	$F_{1,3}$	$F_{2,3}$	$F_{1,2,3}$
レポート課題 A	precision	0.33	**1.00**	0.66	0.66	**1.00**	0.33	**1.00**
	recall	0.08	0.16	0.18	0.22	0.30	0.14	**0.37**
	F 値	0.13	0.28	0.28	0.33	0.46	0.20	**0.54**
レポート課題 B	precision	0.33	**0.66**	0.50	0.16	0.50	**0.66**	**0.66**
	recall	0.22	0.36	0.30	0.12	0.37	**0.44**	**0.44**
	F 値	0.26	0.47	0.37	0.14	0.42	**0.53**	**0.53**
レポート課題 C	precision	**0.50**	0.00	**0.50**	**0.50**	**0.50**	**0.50**	**0.50**
	recall	0.11	0.00	0.14	0.14	**0.83**	0.07	0.10
	F 値	0.18	0.00	**0.22**	**0.22**	0.14	0.12	0.16
レポート課題 D	precision	0.33	1.00	0.66	**1.00**	**1.00**	0.66	**1.00**
	recall	0.12	0.37	0.25	**0.42**	0.37	0.25	**0.42**
	F 値	0.18	0.54	0.36	**0.60**	0.54	0.12	**0.60**

メールによる提出を行っていたためである．

4.2　短文レポート群可視化システム

図 3 は短文レポート評価画面の実行図である．短文レポート評価画面を 4 つに分割して説明する．

図 3 中の①と②は，短文レポートのリストである．①側のリストは，正答らしい短文レポートが並んでいる．反対に②側のリストは誤答らしい短文レポートが並んでいる．各短文レポートの初期状態は，前述の誤答判定機構により，決定され，誤答と判定された短文レポートは②側のリストに追加され，誤答と判定されなかった短文レポートは，①側のリストに追加される．③は 3D グラフである．グラフ上には，三色の点がプロットされている．緑色の点は正答のリストである①に並んでいる短文レポートを指し，赤色の点は誤答のリストである 2 に並んでいる短文レポートを指している．また，青色の点がグラフ上で 1 つだけあり，これは，現在選択中の短文レポートである．③のグラフは Plotly.js[†2] によって実装を行っており，評価者により，自由に回転，拡大／縮小というインタラクティブな操作が可能である．④は，現在選択中の短文レポートの文が表示されている．

リスト上に並ぶ①②と③のグラフ時の点は相互に繋がっており，例えば，③上の点をクリックした際，① ②いずれかのリスト上に存在する点が表す短文レポートの背景色が青色に変更され，①②のリスト上の短文レポートが存在する位置までリストがスクロールされる．

逆に，① ②リスト上の短文レポートをクリックした際，③のグラフ上の対応する点がアニメーションとともに，青色へと変更される．上記のどちらの場合においても，④の文は，対応する短文レポートの文へと変更される．

評価者は，②の誤答と判定された解答を中心に，3D グラフの操作を行いながら，短文レポートの評価を行う．短文レポートの正答誤答の相互移動は，①②のリスト上の各要素が持つ（正答へ／誤答へ）ボタンをクリックすることで移動する．評価を終える際は，画面下部の確定ボタンをクリックする．このボタンをクリックすることで，データベースへの結果の格納とへのメールによる結果の送信が行われる．

5　誤答判定についての評価実験

5.1　実験内容

3 章で提案した 3 つの特徴量を用い，誤答予測システムに最も適した組み合わせを比較実験を行う．実験に用いるデータは，本学の学部三年生向けの講義である自然言語処理の過去 10 年分の短文レポートより，評価済みデータより 4 件（レポート課題 A，B，C，D）を用いる．また，最も誤答を判定できた組み合わせについて，3D グラフにプロットを行い，その考察を行う．

†2　https://plot.ly/javascript/

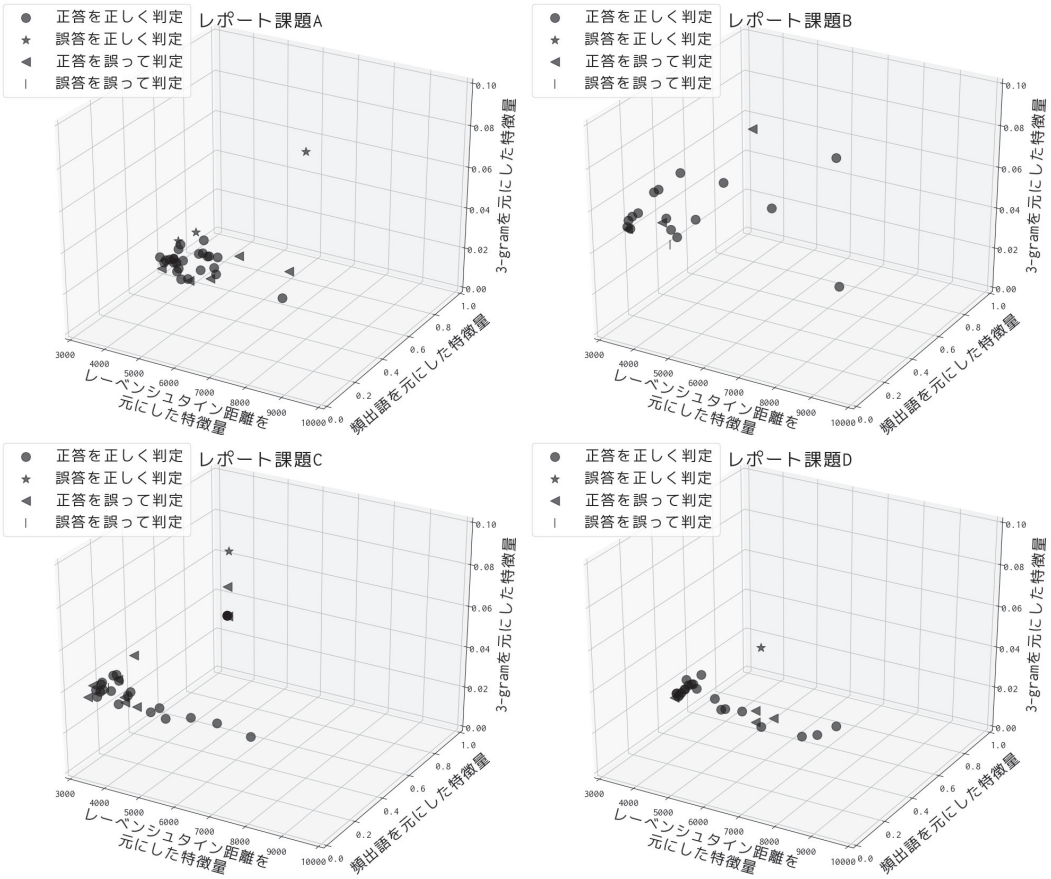

図 4　$F_{1,2,3}$ を用いたグラフ

5.2　実験結果と考察

4 つの短文レポート課題 A,B,C,D に対し，誤答判定機構を用いて，比較評価実験を行った．表 2 は実験の結果を表す．表 1 において，各行はそれぞれのレポート課題を表し，各列は各特徴量の組み合わせを表す．

比較実験の結果について，考察を行う．まず，レポート課題 C において $F2$ の値が 0 であることについて，これは，誤答判定機構により，全ての解答が正答であると判定されたためである．比較実験の結果，$F_{1,2,3}$ が最も良い特徴量の組み合わせであることがわかった．

図 4 は $F_{1,2,3}$ の 3D グラフである．また，グラフ上で，正しく判定されたレポートのうち，正答が正答と判定されたレポートは，“○” マークで，誤答が誤答と判定されたレポートは，“∗” マークでプロットされている．そして，また，グラフ上で，誤って判定されたレポートのうち，正答が誤答と判定されたレポートは，“△” マークで，誤答が正答と判定されたレポートは，“|” マークでプロットされている．

表 2 と図 4 を見ると，F_2 が，判定能力が最も高い特徴量であるといえる．しかし，レポート課題 C における $F2$ の値が 0 であり，単純に頻出語のみを用いた特徴量の組み合わせを用いるべきではないことがわかる．また，レポート全体を通して，より原点に近い位置に正答が集中し，原点より遠い位置に誤答がプロットされることが確認できる．レポート B のグラフは，レポート A,C,D と比較し，グラフ上で短文レポートが散逸している．これは解答の根拠となる講義資料が複数ページに渡って掲載されていたことによる

ものであると判定できる．F_1 が全体を通してあまり誤答の判定に貢献していないことが実験結果よりわかる．

6 おわりに

本稿では，レポート評価支援のため，対象とする短文レポートの特徴より，正答と誤答を判定することを目的に，3 つの特徴量を設計し，3D グラフによるレポートの可視化を用いたレポート評価支援システムの開発を行った．3 つの特徴量の最適な組み合わせにおける実験について，$F_{1,2,3}$ がもっと適した組み合わせであることが分かった．また，F_1 の，判定への貢献が余り大きくないことも分かった．今後の課題として，より誤答の判定に貢献するよう特徴量を改善することと，3D グラフを用いたレポート評価支援システムが，レポート評価者の負担軽減にどの程度繋がるかの実験を行いたい．

$$r_i^{freq} = \frac{1}{\sum_{w_j \in W_i \cap W^{freq}} count(w_j, r_i)} \tag{3}$$

謝辞

本研究の一部は JSPS 科研費 JP15K00422, JP16K00420 の助成を受けたものです．

参考文献

[1] Junbunm, Cha. Jeongwoo, K. S. P.: GRiD: Gathering Rich Data from PubMed Using One-class SVM, *Proc. 2016 Systems,Man,andCybernetics*, IEEE, 2016.

[2] Scholkopf, B. and Platt, J.: Estimating the support of a high-dimensional distribution., *Neural computation*, Vol. 7, No. 13(2001), pp. 1443–1471.

[3] 山添久稔. 大囿忠親. 新谷虎松: 異常検知を用いたレポート評価システムの試作, 電気・電子・情報関係学会東海支部連合大会, 2017(to appear).

[4] 椿本弥生. 柳沢晶義. 赤堀侃司: レポート内容とその評価を可視化する円錐形レポート採点支援マップの開発と評価, 日本教育工学会論文誌, (2007), pp. 317–326.

[5] 石岡恒憲. 亀田雅之. 劉東岳: 人工知能を利用した短答式記述採点支援システムの開発, 研究報告自然言語処理, 2016.