アイトラッキングを利用した，次世代の要件定義書レビュー評価手法
Next-generation review evaluation method of requirement definition document using eye tracking data

斉藤 功樹[†]，土肥 拓生[‡]
Koki Saito, Takuo Doi

[†]日本ユニシス株式会社，[‡]株式会社レベルファイル
Nihon Unisys, Ltd, Level Five Co., Ltd
koki.saito@unisys.co.jp

Abstract

We hereby propose two methods to evaluate the requirement definition document (RDD) review in a page unit using eye tracking data. Eye tracking data were collected from 19 participants as they reviewed the RDD that intentionally included defects and sections that were difficult to read. The first method involved an evaluation of the review quality of the reviewers. We built a classification model to classify individuals who could not identify the included defects using weighted Support Vector Machine and the eye tracking data. An accuracy of approximately 81% was achieved. The second method entailed an evaluation of the difficulty involved in reading the RDD. It was determined that a strong positive correlation is suggested between the re-read count calculated using the eye tracking data and the reading difficulty level.

Keywords — Eye Tracking, Review, Requirement Definition Document, Machine Learning

1. はじめに

ソフトウェア開発において，上流工程における仕様書や設計書の品質が，後工程の成果物の品質へも影響を及ぼすため，上流工程における仕様書の品質を担保することが重要である．上流工程における要求仕様書や要件定義書の質を高める一般的な手法としてレビューが挙げられ，様々なレビュー手法が存在する．しかし，不具合検出率などのレビュー品質に及ぼす影響は，レビュー手法による違いよりも個人差のほうが大きい[1]．同じ個人においても，時間的な制約や集中度合いなどにより，レビューごとにレビュー品質が異なることがあり，個々のレビューの評価が難しい．

レビュー品質の定量的な評価は，レビュー実施率やレビューでの指摘を基にした不具合検出率が用いられるが，それらの指標だけでは，レビュアーの能力によるのかレビュー対象文書の品質によるのか判別できず，レビュー品質を正確に評価することができない．また，不具合に繋がらない文章自体の読みづらさを評価する指標も存在しない．したがって，レビュー実施率や不具合検出率によらずに，個々のレビューの評価やレビュー文書の読みづらさが評価できれば，ソフトウェア開発の品質向上において非常に有効である．

レビュー時におけるレビュー品質を評価するためにはレビュー時の集中力やレビュー対象文書に対する理解度などが影響すると考えられる．多くの研究で，より注意が必要になればなるほど，瞬きが少なくなることが結論付けられおり[2]，JINS では，瞬きを用いて人の集中力を可視化するアルゴリズムを開発している[3]．英語の理解度と視線の関係についても研究されており，英語の理解度に影響を与える視線の特徴を抽出する研究[4]や視線を用いた英語スキルの推定の研究[5]などが行われている．また，文章が難しいと感じた場合は，読む速度が遅くなる，何度も読み返す，同じ場所を注視し続けるといった傾向がある[7]．視線情報を活用することで，レビュー時のレビュー品質の評価やレビュー文書自体の読みづらさの評価ができると考える．

本稿では，要件定義書レビューを対象として，レビュー時の視線情報を用いて，個々のレビューを 1 ページ単位で評価する 2 つの手法を提案する．1 つ目はレビュー時のレビュー品質評価手法である．適切にレビューできる人，あるいは，適切にレビューできるだけ集中している際には，そのレビューの仕方に共通点があると考え，レビュー時の視線情報を機械学習により適切なレビューがなされているかを判別するモデルを構築する手法である．

2 つ目はレビュー文書の読みづらさを自動的に評価する手法である．読みづらいレビュー文書をレビューしている際の視線は，読みやすいレビュー文書をレビューしている場合と比較して，相違があるとの仮説に立ち，レビュー時の視線からもレビュー文書を評価する手法である．

2. 提案手法

2.1. レビュー時のレビュー品質評価

要件定義書レビュー時の視線情報を用いて，1 ページ単位でレビュー時のレビュー品質を評価する手法を提案する．アイトラッカを用いてレビュー時の視線情報を取得し，取得した視線情報を基にして 1 ページ単位で説明変数を算出する．アイトラッカにより計測可能なデータは多数あるが，レビュー品質と相関のないデータも多数あると考えられるため，まずは，説明変数の次元削減を行い，レビュー品質に影響を及ぼしている説明変数を抽出する．抽出した説明変数を基にして，機械学習を用いることでレビュー品質を評価するモデルを構築する．

レビュー時に適切にレビューできない人を検出できれば，品質の悪いレビューが特定でき，要件定義書の品質向上に寄与できるため，本稿では，適切にレビューできない人を精度よく分類できるモデルを構築することを目的とする．

2.1.1. 説明変数の次元削減

説明変数は, Bixler らの研究[7]で定義されていた 46 個の変数に，単位時間あたりの瞬きの回数を追加した 47 個の説明変数をベースとする．47 個の説明変数を表 1 に示す．

47 個の説明変数において，レビュー品質評価において重要である変数を特定するために，ランダムフォレスト（以降，RF）を行い，重要度が上位の説明変数を用いてモデルを構築する．

2.1.2. アルゴリズムの選定と不均衡データ対応

2.1.2 で次元削減を行った説明変数を用いて，Support Vector Machine（以降，SVM）にてモデルを構築し，レビュー時のレビュー品質評価モデルを構築する．

適切にレビューできる人とできない人の被験者における分布が同一でない不均衡データであることが想定されるため，SVM において重み付けを行い，不均衡データへの対応を行う．

表 1　47 個の説明変数

説明変数			詳細
fixation	duration	minimum	fixationの持続時間の最小
		maximum	fixationの持続時間の最大
		mean	fixationの持続時間の平均
		median	fixationの持続時間の中央値
		standard deviation	fixationの持続時間の標準偏差
		skew	fixationの持続時間の歪度
		kurtosis	fixationの持続時間の尖度
		range	fixationの持続時間のレンジ
		dispersion	fixation中の視点の分散
saccade	distance	minimum	saccadeの距離の最小
		maximum	saccadeの距離の最大
		mean	saccadeの距離の平均
		median	saccadeの距離の中央値
		standard deviation	saccadeの距離の標準偏差
		skew	saccadeの距離の歪度
		kurtosis	saccadeの距離の尖度
		range	saccadeの距離のレンジ
	duration	minimum	saccadeの持続時間の最小
		maximum	saccadeの持続時間の最大
		mean	saccadeの持続時間の平均
		median	saccadeの持続時間の中央値
		standard deviation	saccadeの持続時間の標準偏差
		skew	saccadeの持続時間の歪度
		kurtosis	saccadeの持続時間の尖度
		range	saccadeの持続時間のレンジ
	angle	minimum	saccadeの角度の最小
		maximum	saccadeの角度の最大
		mean	saccadeの角度の平均
		median	saccadeの角度の中央値
		standard deviation	saccadeの角度の標準偏差
		skew	saccadeの角度の歪度
		kurtosis	saccadeの角度の尖度
		range	saccadeの角度のレンジ
	other	count	saccadeの回数
		proportion of horizontal	水平なsaccadeの割合 (水平軸から30度未満の角度)
pupil	diameter (z-score で標準化)	minimum	pupilのdiameterの最小
		maximum	pupilのdiameterの最大
		mean	pupilのdiameterの平均
		median	pupilのdiameterの中央値
		standard deviation	pupilのdiameterの標準偏差
		skew	pupilのdiameterの歪度
		kurtosis	pupilのdiameterの尖度
		range	pupilのdiameterのレンジ
blilnk		count	blinkの回数
		count ratio	blinkの回数の割合
		time	blinkの時間の割合
other		fixation duration /saccade duration ratio	fixationの持続時間/ saccadeの持続時間の比率

2.2. レビュー文書の読みづらさ評価

要件定義書レビュー時の視線情報を用いて，1 ページ単位でレビュー文書の読みづらさを評価する手法を提案する．レビュー時のレビュー品質評価と同様に，アイトラッカを用いてレビュー時の視線情報を取得し，取得した視線情報を基にして 1 ページ単位で読み返しの回数を算出する．算出した読み返しの回数と各ページの単語数を基にして，各ページの単語数ごとの読み返しの回数を算出する．単語数ごとの読み返しの回数

を読みづらさの指標として用い，レビュー文書の読みづらさを評価する．

2.2.1. 読み返し判定

大社らの研究[6]で定義されていた読み返しの判定式を用いた．図 1 のように，ある n 個目の fixation をF_n，その座標を(X_n, Y_n)，その前後の fixation とのなす角度をA_nとする．読み返しは段落をまたぐものもあるが，本稿では大社らの研究と同様に単語，イディオム程度を繰り返す読むことを読み返しとする．図 2 のフローチャートに従い，読み返しを判定する．H，W，Kはそれぞれ閾値を指す．

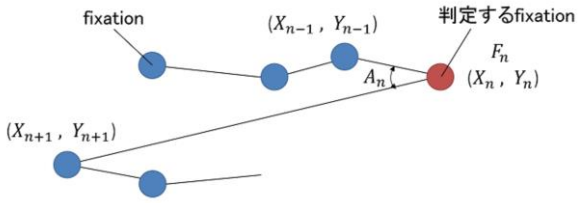

図 1　読み返し判定の際の fixation の様子（出典：大社・Kunze・Augereau・黄瀬（2015））

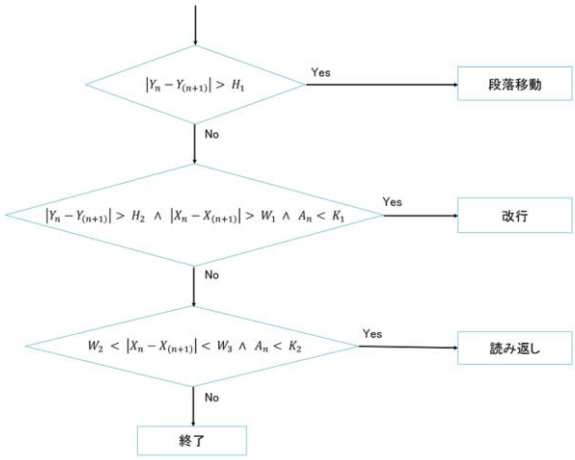

図 2 読み返し判定のフローチャート（出典：大社・Kunze・Augereau・黄瀬（2015））

2.2.2. 読みづらさ評価

2.2.1 にて判定した読み返しを基に，1 ページ単位で単語数ごとの読み返しの回数を算出して，読みづらさ評価に用いる．単語数ごとの読み返しの回数とレビュー文書の読みづらさの相関関係を調査し，読みづらさが増したページにおいては単語数ごとの読み返しの回数が増加することを確認する．

3.　実験

3.1.　実験条件

レビュー対象文書は，弊社で実際に使用された 3 種類の要件定義書を基に抜粋して，概要・機能要件・非機能要件の 3 ページ構成とし，サンプル文書を 2 つ加え，計 11 ページとする．それぞれの要件定義書の概要を表 2 に示す．それぞれの要件定義書の各ページに後工程での障害に繋がる欠陥を含ませ，1 ページあたりに欠陥を含む文章は最大 2 個までとし，11 ページ全体で欠陥を含む文章は 16 個とする．それぞれの要件定義書の機能要件において，表形式や箇条書き形式を文章形式に変更し，システム化対象範囲の図を削除することで，概要・非機能要件は変更を加えた機能要件と比べて読みやすく，変更を加えた機能要件は読みづらさが増加するように変更した．

19 名の被験者に対して，ディスプレイ上に要件定義書を提示し，レビュー時の視線をアイトラッカ（GP3 HD 150Hz）にて記録した．被験者には，ディスプレイ上の要件定義書をレビュー後，紙の要件定義書に対して改善点を下線にて記載するよう指示を与えた．途中休憩は挟まず，レビュー時間・形式などの制限を与えなかったが，測定の精度を高めるためになるべく頭部を動かさないように指示を与えた．下線にて記載した改善点が，意図的に含んだ欠陥部分と一致した場合に欠陥を検出できたとし，全 11 ページでの欠陥検出数を総欠陥検出数とする．実験終了後にアンケートを実施した．実験の装置を図 3 に示す．

表 2　実験で使用した要件定義書の概要

要件定義書種別	概要
1	チャージポイントゲートウェイシステム開発について
2	通販システムの基幹システム刷新について
3	踏切定常監視システム構築について

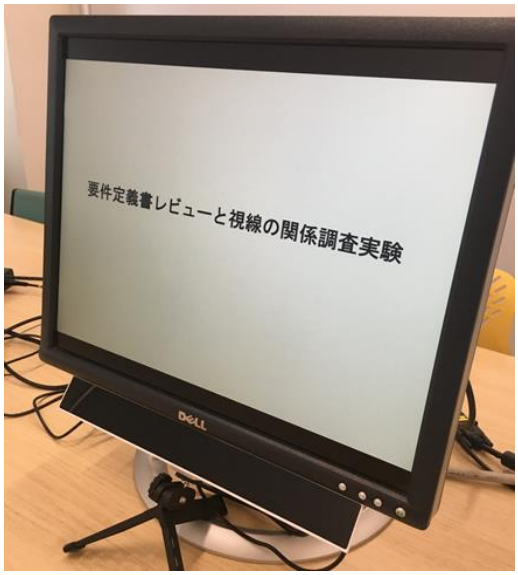

図 3　実験の装置（ディスプレイの下部：アイトラッカ）

3.2. 被験者特性

実験終了後のアンケートでは，年齢・性別の基本情報の他に，実験時の集中度・文書に対する理解度・要件定義書レビュー経験・文書レビュー頻度の被験者特性を取得した．被験者構成の年代は 30 代～50 代であり，年齢の分布を図 4 に示す．要件定義書レビュー経験数は要件定義書レビューを含むプロジェクトに携わった数を示しており，約半数の被験者は要件定義書レビュー経験が全くない．被験者の要件定義書レビュー経験の分布を図 5 に示す．

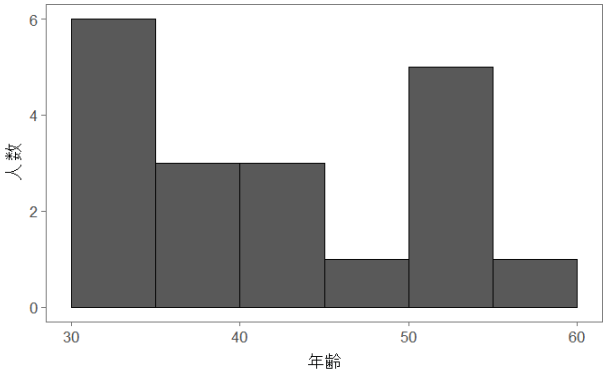

図 4　被験者の年齢分布

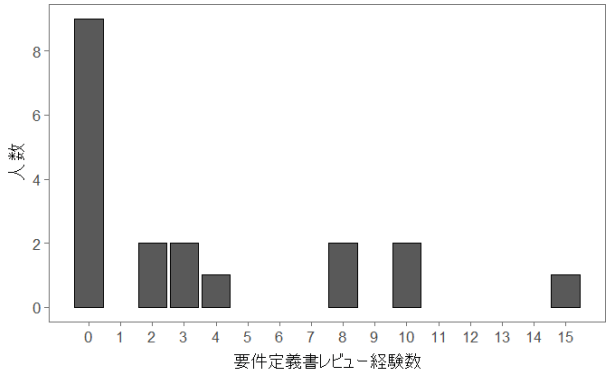

図 5　被験者の要件定義書レビュー経験分布

4.　モデルの作成・評価

4.1. レビュー時のレビュー品質評価

総欠陥検出数の被験者分布を図 6 に示す．レビュー品質評価では，総欠陥検出数が 1 以下の被験者を欠陥を検出できない人として定義し，総欠陥検出数が 1 以下の被験者を low グループ，2 以上を high グループとして，low グループを精度よく分類できるモデルを構築する．

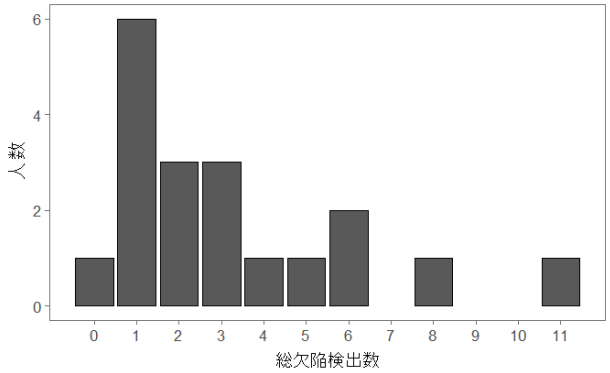

図 6　総欠陥検出数の被験者分布

4.1.1. データの前処理

被験者のページごとの視線情報において，適切に視線情報が取得できていないページの除外を行った．アイトラッカから得られる fixation が有効かどうかの情報を基にして，ページごとの有効な fixation の割合を算出し，スミルノフ・グラブス検定を用いた．有効な fixation の割合のヒストグラムを図 7 に示す．判定の結果, 図において有効な fixation の割合が 0.6 未満の 4 ページ分を外れ値として除外した．

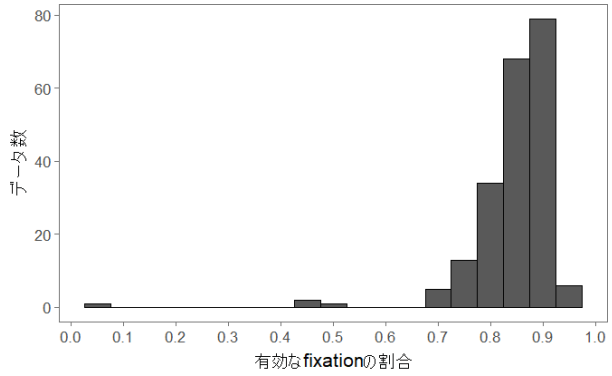

図 7　ページごとの有効な fixation の割合のヒストグラム

4.1.2. 説明変数の次元削減

ページ単位で算出した 47 の説明変数を基にして，RF における重要度が上位の説明変数を用いて分類を行った．RF を複数回実行した場合において，常に 9 つの説明変数の重要度が上位にあったため，上位 9 つの説明変数を採用した．

主成分分析を行った累積寄与度の 80%以上の主成分とアンケート結果を用いて分類を行った結果を比較対象として用いる．それぞれの説明変数を用いて，Leave-one-out cross validation を行い，Recall，Accuracy を算出した結果を表 3 に示す．Recall，Accuracy は leave-one-out cross validation を 19 人分行った際の平均値である．

RF における重要度が上位 9 つの説明変数を用いて分類モデルを構築した場合の Recall, Accuracy が最も高く，レビュー品質評価においては 9 つの説明変数が重要であることがわかった．

表 3　説明変数ごとの評価指標

アルゴリズム	評価指標	項目	データ数	説明変数			
				アンケート	47個	主成分	RF_上位9
SVM	Recall	high	131	77.3%	71.2%	72.0%	**85.4%**
		low	74	55.8%	10.9%	16.2%	**72.2%**
	Accuracy	all	205	69.4%	49.0%	51.4%	**80.5%**

4.1.3. 提案手法と他アルゴリズムの比較

RF における重要度が上位 9 つの説明変数を用いて SVM にて分類を行った．比較対象として，SVM 以外のアルゴリズムを用いた．それぞれのアルゴリズムにおいて，Recall，Precision，Accuracy を算出した結果を表 4 に示す．test データとして low，high それぞれ 1 名ずつ，train データとして残りの全データとし，すべての組み合わせにおおける Recall，Precision，Accuracy の平均値を算出した．

SVM（重み付け）では，データ数が少ない low において最も Recall が高くなり，SVM（重み付け）を用いることで low のグループを最も精度高く分類できる．

表 4　アルゴリズムごとの評価指標

評価指標	項目	データ数	RF	SVM（重み付け）	DT	kNN	NNET
Recall	high	131	87.0%	75.5%	**88.5%**	72.7%	84.1%
	low	74	71.4%	**81.1%**	77.7%	62.6%	71.3%
Precision	high	131	77.6%	80.0%	**81.8%**	65.5%	76.9%
	low	74	87.9%	82.0%	**90.1%**	75.6%	84.8%
Accuracy	all	205	79.1%	78.3%	**82.9%**	67.7%	77.7%

4.1.4. モデルの評価・考察

以上の結果より RF の重要度上位 9 つの説明変数を採用し，SVM（重み付け）を用いることで，low グループを最も精度高く分類できることがわかった．

SVM（重み付け）を用いて，train データを変化させて算出した Accuracy の学習曲線を図 8 に示す．train データを変化させた場合の train データ・test データにおける low と high グループの被験者の割合は表 5 のとおりである．train データが増加するにつれて，train データと test データでの Accuracy が近い値に収束しており，過学習は発生していないと考えられる．しかし，Accuracy が完全には収束しておらず，train データを増やすことで更なる精度向上が見込める．

RF の重要度上位の 9 つの説明変数において，最も重要度が高い説明変数は瞬きの時間の割合であり，次いで単位時間あたりの瞬きの回数であった．構築した分類モデルにおいて，瞬きが少ないまたは多い場合に high に分類される傾向があった．注意が必要なタスクでは瞬きが少なくなるといわれており，注意深く文章を読んでいる間は瞬きが少なくなると予測される．high のグループの人は，欠陥を指摘できているため注意深く読んでいると考えられるが，瞬きの回数の割合が多い場合にも high のグループが存在している．ページごとの瞬きの時間の割合のヒストグラムを図 9 に示す．既存研究からは，欠陥を検出できる特徴として瞬きが少ないことが考えられたが，瞬きが多いことも特徴の一つであることがわかった．

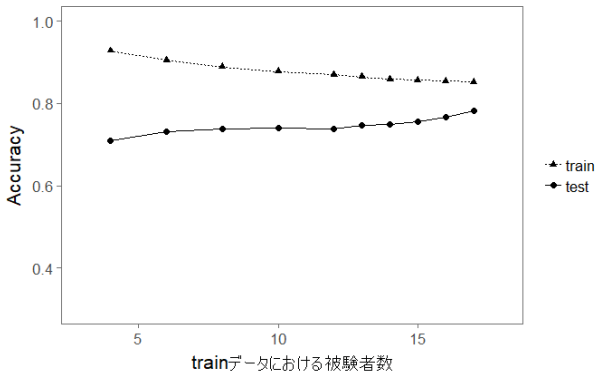

図 8　学習曲線

表 5　学習曲線における被験者数

		被験者数									
trainデータ	high	2	3	4	5	6	7	8	9	10	11
	low	2	3	4	5	6	6	6	6	6	6
	計	4	6	8	10	12	13	14	15	16	17
testデータ	high	10	9	8	7	6	5	4	3	2	1
	low	5	4	3	2	1	1	1	1	1	1
	計	15	13	11	9	7	6	5	4	3	2

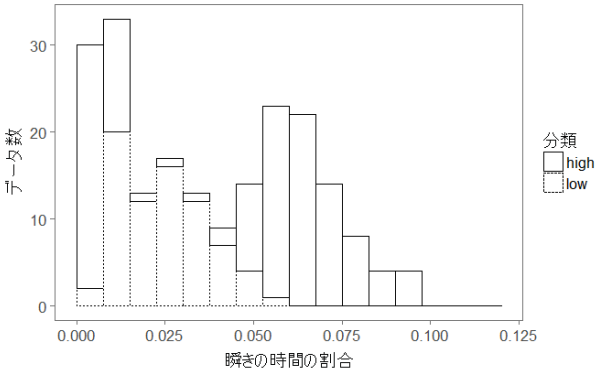

図 9　ページごとの瞬きの時間の割合のヒストグラム

4.2. レビュー文書の読みづらさ評価

読みづらさ評価には，レビューにおいて流し読みをせずに文章を読んでいる人を対象とする必要があるため，欠陥の検出数が多い人を対象とする．それぞれの要件定義書において，読みづらさの順位尺度を以下のように定義する．読みづらさを増すように変更を加えた機能要件のページの読みづらさの順位尺度を 2，変更を加えていない概要と非機能要件のページの順位尺度をそれぞれ 1 とした．

それぞれの要件定義書の各ページの単語数ごとの読み返しの回数を基に，読みづらさの順位尺度との相関関係を調査した．比較対象の指標として読み返しの総回数，レビュー時間，単語数ごとのレビュー時間を用いた．全ての要件定義書における概要・非機能要件と機能要件それぞれの各指標の平均値と平均値の増加率を表 6 に，それぞれの要件定義書における読みづらさと各指標の相関係数を表 7 に示す．要件定義書の各ページにおける単語数ごとの読み返しの回数の平均値を表 8 に，箱ひげ図を図 10 に示す．1-*はそれぞれ要件定義書の種別を，*-1 は概要，*-2 は機能要件，*-3 は非機能要件のページを示す．相関係数と平均値の算出には，箱ひげ図における外れ値は含めなかった．

読みづらさが増した機能要件のページにおいては，概要・非機能要件のページと比較して，単語数ごとの読み返しの回数の増加率が最も高い結果となり，単語数ごとの読み返しの回数が読みづらさの増加を顕著に示すことがわかった．それぞれの要件定義書においては，要件定義書種別 1 と 3 では単語数ごとの読み返しの回数と読みづらさと相関が高いが，種別 2 では相関がみられなかった．種別 2 においては，機能要件の記載内容が一般的な web システムのログインなどの機能であり，他の要件定義書と比較して内容自体に読みやすい要素があったためと考えられる．文章の内容自体を考慮する必要があるが，単語数ごとの読み返しの回数を用いることで，文章の構造変化による読みづらさを評価できることがわかった．

単語数ごとのレビュー時間においても単語数ごとのレビュー時間と同様の傾向を示すが，要件定義書種別 3 では単語数ごとの読み返しの回数と比べて相関が弱くなっている．レビュー時間は，被験者がレビュー文書を読んでいる以外の思考の時間や文章全体の読み返しなどの時間も含むため，文章の読みづらさ評価の指標としては適さないと考えられる．

表 6　読みづらさと各指標の平均と増加率

読みづらさの順位尺度	読み返しの総回数		単語数ごとの読み返しの回数		レビュー時間		単語数ごとのレビュー時間	
	平均	増加率	平均	増加率	平均	増加率	平均	増加率
1（概要・非機能要件）	28.1852	-	0.0545	-	166.744	-	0.3187	-
2（機能要件）	37.1429	32%	0.0811	**49%**	184.854	11%	0.4025	26%

表 7　読みづらさと各指標の相関係数

要件定義書種別		読み返しの総回数	単語数ごとの読み返しの回数	レビュー時間	単語数ごとのレビュー時間
1	Spearmanの相関係数	0.3174	0.5071	0.3118	**0.5345**
	p値	0.2907	0.0769	0.2997	**0.0599**
2	Spearmanの相関係数	0.0186	0.0925	-0.0423	-0.1268
	p値	0.9497	0.7533	0.8910	0.6798
3	Spearmanの相関係数	0.6325	**0.7076**	0.3563	0.5345
	p値	0.0152	**0.0046**	0.2320	0.0599

表 8　単語数ごとの読み返しの回数の平均

読みづらさの順位尺度	2（機能要件）			1（概要・非機能要件）					
要件定義書ページ種別	1-2	2-2	3-2	1-1	1-3	2-1	2-3	3-1	3-3
単語数ごとの	0.0932	0.0656	0.0855	0.0430	0.0527	0.0673	0.0591	0.0561	0.0486
読み返しの回数の平均	0.0811			0.0545					

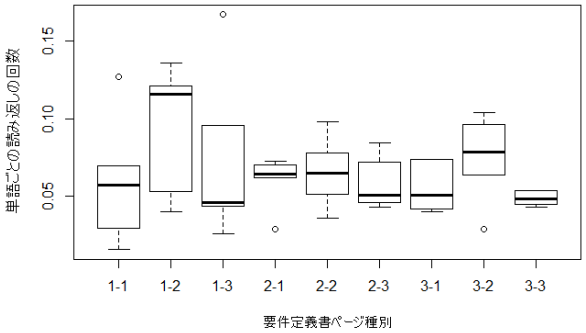

図 10　単語数ごとの読み返しの回数の箱ひげ図

5.　まとめ

本稿では，視線情報を利用して要件定義書レビューを 1 ページ単位で評価する 2 つの手法を提案した．

レビュー品質評価では，9 つの視線情報を用いて欠陥を検出できていない人のレビューを SVM（重み付け）にて約 81%の精度で分類できるモデルが構築できた．分類において重要な視線情報は 9 つであったが，特に瞬き時間の割合と単位時間当たりの瞬きの回数が重要であった．欠陥を検出できる読み方の視線の特徴の一つとして，従来の研究からは瞬きが少ないことが考えられたが，瞬きが少ないだけではなく瞬きが多いということも重要な特徴であることがわかった．

レビュー文書の読みづらさ評価では，文章の構造を変更することで読みづらさが増したページに対しては，単語数あたりの読み返しの回数が増加し，読みづらさと読み返しの回数に最大 0.71 の相関係数を示し，正の相関関係があることがわかった．文書の内容を考慮する必要があるが，読み返しの回数を用いることで，レビュー対象文書の読みづらさが評価できる．

参考文献

[1] H. Uwano, M. Nakamura, A. Monden, and K. Matsumoto, (2006) “Analyzing Individual Performance of Source Code Review Using Reviewers’ Eye Movement”, in Proceedings of the 2006 Symposium on Eye Tracking Research & Applications, pp. 133-140

[2] H. Ledger, (2013) “The effect cognitive load has on eye blinking”, Plymouth Student Sci., Vol. 6, No. 1, pp. 206-223

[3] Y. Uema and K. Inoue, (2017) “JINS MEME algorithm for estimation and tracking of concentration of users”, UbiComp/ISWC 2017 - Adjun. Proc. 2017 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Proc. 2017 ACM Int. Symp. Wearable Comput., pp. 297-300

[4] A. Okoso, T. Toyama, K. Kunze, J. Folz, M. Liwicki, and K. Kise, (2015) “Towards Extraction of Subjective Reading Incomprehension: Analysis of Eye Gaze Features”, Chi Ea, pp. 1325-1330

[5] O. Augereau, K. Kunze, H. Fujiyoshi, and K. Kise, (2016) “Estimation of english skill with a mobile eye tracker”, Proc. 2016 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Adjun. - UbiComp ’16, pp. 1777-1781

[6] 大社綾乃, K. Kunze, O. Augereau, 黄瀬浩一, (2015) “学習補助のための視線情報に基づく文書アノテーション”, 電子情報通信学会技術研究報告, PRMU2015-30, Vol. 115, No. 24, pp. 161-166

[7] R. Bixler and S. D’Mello, (2015) “Automatic Gaze-Based Detection of Mind Wandering with Metacognitive Awareness”, in User Modeling, Adaptation and Personalization, pp. 31-43