AI（機械学習）を利用した文書検索による障害報告書と開発ノウハウの活用度向上

Improved use of trouble reports and development know-how through document search engine using AI (machine learning)

田淵　秀之

hideyuki.tabuchi@mizuho-ir.co.jp

みずほ情報総研株式会社　品質管理部

発表要旨：

レガシーシステムで培ってきた品質を確保するための開発ノウハウは、デジタル化の潮流など新しい技術が注目される現在も重要と考え、品質管理部門は、障害報告書を活用した開発ノウハウ可視化に取り組んでいる。この取組みは「経験不足を補う良い活動」といった声がある一方で、開発ノウハウ可視化作業は経験のあるベテラン社員が相応の工数を要して作成することから、開発ノウハウ数がなかなか伸びないという問題がある。また、開発ノウハウへのアクセスが少なく活用度が低いという問題もある（テキストで公開しているため検索できないという背景もあり）。

これら問題に対して、AI（機械学習）を利用した文書検索エンジンで解決できないか、試作品を開発し効果検証した。問題解決のアプローチは、可視化数がなかなか伸びない問題に対しては、開発ノウハウで紹介している障害事例と、同水準の件数とバリエーションの障害報告書を文書検索エンジンで検索できれば、障害の注意喚起機能を代替できると考えた。ノウハウへのアクセスが少ない問題に対しては、ネット検索のような利便性の高い検索エンジンを作ることができれば開発ノウハウへのアクセス数が伸びると考えた。本プログラムでは、効果検証の結果、文書検索エンジンの開発で苦労／工夫したことを紹介する。

キーワード：

品質管理,障害報告書・開発ノウハウの有効活用，AI（機械学習）

想定している聴衆

・障害報告書の活用が必要と考えている人

・開発ノウハウの可視化、浸透を推進している人

・ＡＩ（機械学習）を活用した品質管理業務の改善を検討している人

発表者の紹介（全角１００文字）：

ＱＭＳの導入・推進に長く従事し３社で登録認定に関わる。一方、開発部署で大規模システムのＰＭを経験。現在、プロジェクト管理研修や開発ノウハウ可視化など人材育成業務に従事。ＳＱｕＢＯＫ（Ｖ１）を共著。

AI（機械学習）を利用した文書検索による障害報告書と開発ノウハウの活用度向上

2020.09.11

みずほ情報総研　株式会社
品質管理部　　田淵秀之
E−mail：hideyuki.tabuchi@mizuho-ir.co.jp

目次

1. 背景と問題　①背景

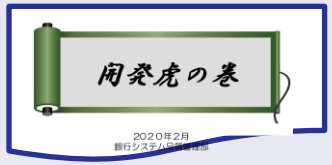
デジタル化の潮流で新技術への関心が高まる現在においても

レガシーシステムで培ってきた品質を確保するための開発ノウハウの継承は重要

⇐ ベテラン社員の退職

障害報告書を活用した開発ノウハウ可視化の取組

開発ノウハウ

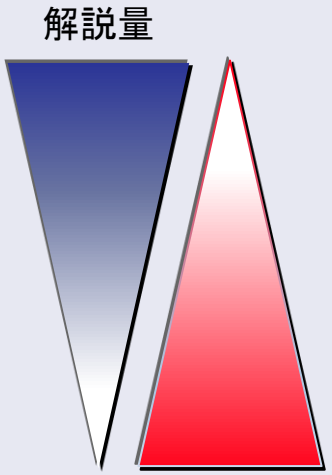


技術的な解説をしっかり伝える必要がある時に利用

ケーススタディ集

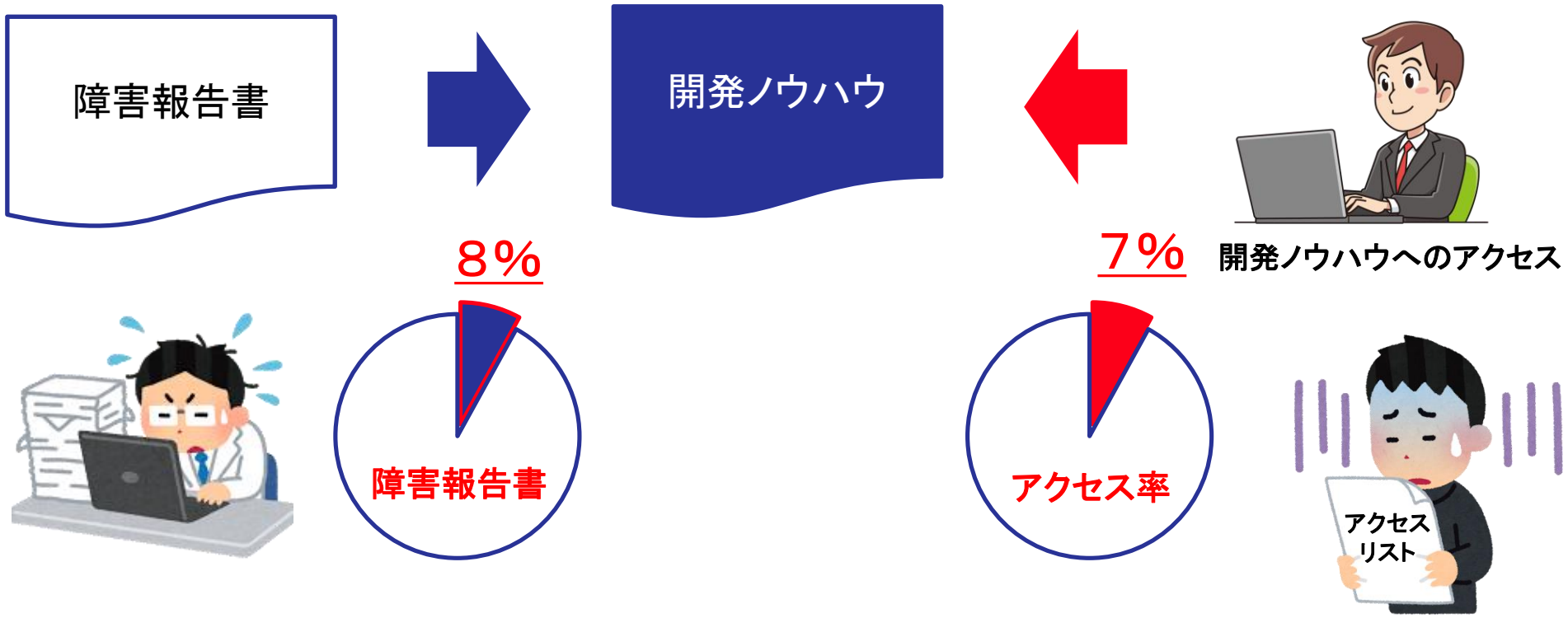
障害発生の原因（問題点）だけでなく、解決策の紹介が重要な時に利用

技術的な解説より、障害リスク事例の紹介が重要な時に利用



経験不足を補う良い活動

高評価



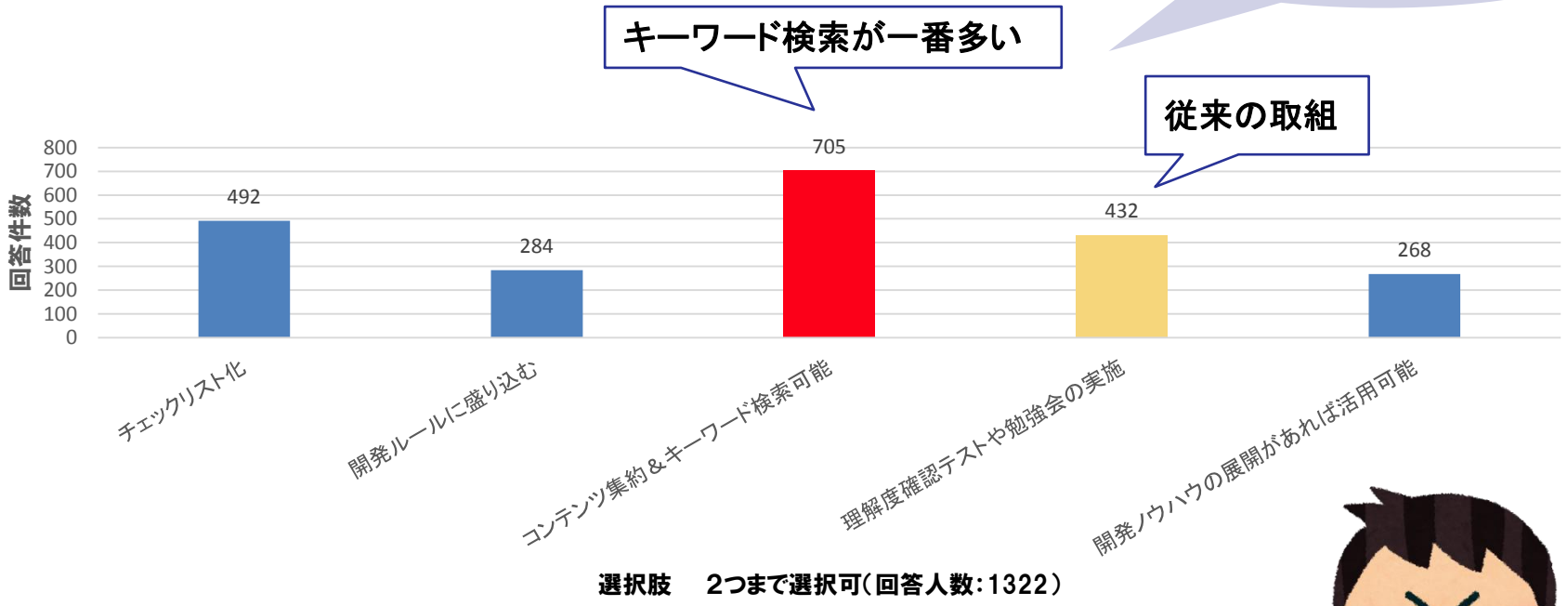
開発ノウハウ可視化は工数を要するため多くは発信できず、可視化した開発ノウハウは障害報告書の8%程度に留まっている

開発ノウハウへのアクセスが少なく活用度が低い
（発信直後に社員の7%程度がアクセスする程度）

アクセス数が少ない実態を認識して、、、

「開発ノウハウを活用する方法」をアンケートでヒアリング！

フリーコメントで利便性向上に関する要望が24件あり



選択肢　2つまで選択可（回答人数：1322）

使い勝手の良い検索エンジンを作ろう！

2. 研究概要　①問題解決アプローチ

AI（機械学習）を利用した文書検索エンジン（以下、AI検索エンジン）の試作品を開発し、問題解決が可能か検証する

問題点	課題設定	期待効果	課題設定の根拠
提供側の問題	**有用性**		
可視化した**開発ノウハウは障害報告書の8%程度**に留まっている	開発ノウハウで紹介している障害事例と、同水準の件数とバリエーションの障害報告書を検索できる	開発ノウハウ可視化を待つことなく、開発ノウハウと同水準の障害リスクを把握することができる	**開発ノウハウは、障害リスクの注意喚起という役割も大きい**（「あるあるトラップ」はそれが主目的）
利用者側の問題	**利便性**		
開発ノウハウへのアクセスが少なく活用度が低い	ネット検索のように、複合検索ができ、検索語と適合した文書が上位に検索される	ネット検索のように、探したいものが直ぐに見つかる便利な検索エンジンなら、障害報告書や開発ノウハウが検索される	現在、障害報告書や開発ノウハウ（PPT/　Word)をNotesで公開。**検索性の悪さが、低アクセスの原因の1つ**と想定

2. 研究概要　②類義語検索による多面的な検索機能

Word2Vecを利用した類義語検索による多面的な検索機能（問題1への対応）

Word2Vec

Word2Vecモデル

学習

近くにいる単語を抽出して重み付け

類義語検索

障害報告書

nn文書
mmm単語（名詞、動詞、形容詞）

障害番号：20200001
タイトル　：マスタ作成処理の文字コード変換エラー

文字コード変換ツールに以下の制約があるため、変換範囲外の文字は変換エラーとなり文字化けするもの。バリナリ形式の文字コードを変換する場合はSJIS全ての文字コード範囲を変換可能だが、、

検索語　：　文字コード

0.71 文字コード変換
0.65 規格
0.61 SJIS
0.55 変換範囲外
0.53 文字化け
0.52 外字

類似度の高い順に出力

検索語　：　指定した検索語　＋　検索語の類義語

（例）　“文字コード”　＋　“文字コード変換，規格，SJIS，変換範囲外”

追加する類義語の個数を指定可能

ベクトル空間モデルを利用した検索

単語や文書をベクトル化して、ベクトルの向きが近いものが「似た」単語、文書になるという特性を使って、検索文書と似た文書を特定する

＜単語ベクトル＞

単語a　(1,2)　単語b　(2,1)

単語c　(8,2)　単語d　(2,8)

＜類似度（ベクトルの向き）＞

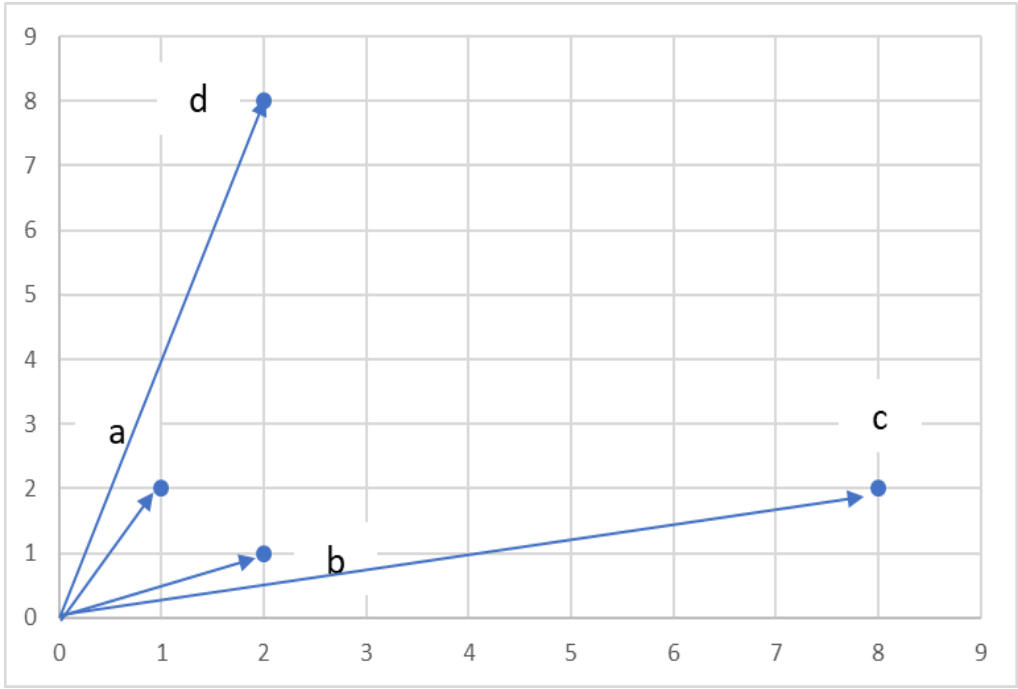
・単語aと単語dは似ている

・単語dと単語cは似ていない

・単語aと単語bは微妙

＜類似度の計算方法：コサイン類似度＞

$$\cos\theta = \vec{a}\cdot\vec{b} \ / \ |a| \ |b|$$



2. 研究概要　③検索文書と適合度の高い順に検索結果出力2/2

ベクトル空間モデルを利用し、検索文書と適合度の高い順で検索結果を出力（問題2への対応）

TF－IDFによる文書のベクトル化

- ✓ 文書中の単語出現頻度（TF）と、単語が出現する文書数の全文書数割合（IDF：希少価値度）を掛け合わせたTF－IDFで単語を数値化し、文書をベクトル化する。
- ✓ 各文書において、単語の出現頻度が高く、希少価値の高い単語は、大きな数値になる。

検索文書と障害報告書の類似度を計算し、類似度の高い文書から出力する

工夫

- ✓ 障害報告書全文書の類似度計算をすると時間を要して使い物にならない
- ✓ 転置インデックスを利用して、検索語を使っている障害報告書を特定した上で類似度計算を実施する

障害文書のベクトル化（TF-IDF）

		障害報告書全量に含まれる単語						
		単語１	単語２	単語３	単語４	単語５	・・・・・	単語mmm
障害報告書全量	障害1	0.1	0.1	0	0.1	0.1	・・・・・	0.1
	障害2	0.3	0	0	0.3	0.2	・・・・・	0.1
	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・ ・ ・	・・・・・	・ ・ ・
	障害nn	0	0.2	0.3	0	0.1	・・・・・	0.1
	単語1～単語mmmが１語ずつの文書	0.1	0.1	0.1	0.1	0.05	・・・・・	0.07
検索文書	単語３,単語５	0	0	0.1	0	0.05	0	0

mmm次元のベクトル表現
ほとんど０でスカスカ

転置インデックス

単語1	障害1、障害2
単語2	障害1、障害70
・ ・ ・	・ ・ ・
単語mmm	障害1、障害2、障害nn

検索文書に対する類似度計算（文書の類似度計算）

								合計	類似度
障害1	0	0	0	0	0.1*0.05	0	0	0.005	3
障害2	0	0	0	0	0.2*0.05	0	0	0.01	2
・	・	・	・	・	・	・	・	・	・
障害nn	0	0	0.3*0.1	0	0.1*0.05	0	0	0.035	1
検索文書（単語３,単語５）	0	0	0.1	0	0.05	0	0		

類似度でソートする

・類似度の高い順に出力
・複合検索（複数の検索語）を実現！

2. 研究概要　④専門用語の辞書登録

ベクトル空間モデルによる文書検索の問題点と解決方法

検索文書　：“開発環境と本番環境の違いで発生”

①形態素解析　②名詞、形容詞、動詞のみ抽出

検索語　　：“開発”,”環境”,”本番”,”違い”,”発生”

専門用語の辞書登録
（上記例では、“開発環境” “本番環境”を辞書登録することで解決）

「専門用語自動抽出システム」※の利用

専門用語は、たくさん登録すれば良いというものではない

- ✓“CPU使用率”は、登録すべきか／登録しない方が良いか？
  登録すれば、“CPU”では“CPU使用率”を含む障害報告書は検索されない
- ✓抽出された専門用語から辞書登録したものは17%程度

意図しない障害報告書がたくさん検索されてしまう！

沢山ある単語から、必要な専門用語をどうやって見つけるの？

工夫

専門用語の辞書登録は、悩ましい問題＆選定に多くの時間が必要

※東京大学情報基盤センター図書館電子化部門中川裕志教授および 横浜国立大学環境情報研究院森辰則助教授が共同開発

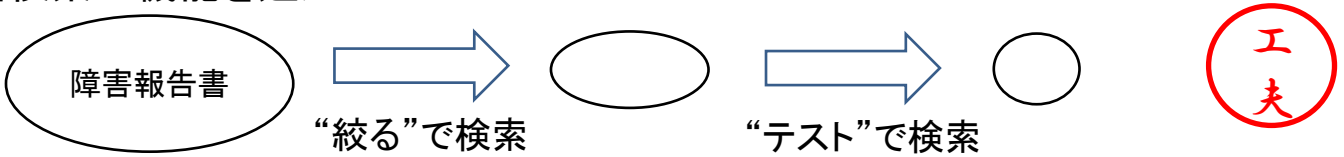
3. 検証結果　①有用性　1/3

AI検索エンジンの有用性を示すため、以下の比較を実施

- 開発ノウハウで紹介してきた障害事例件数
- 障害報告書の文字列検索結果、AI検索結果

開発ノウハウ	開発ノウハウ（障害事例）	検索語	文字列検索	AI検索（類義語含む）
文字コードについて	4件	文字コード	12件	適合　15件 検索　30件 適合最低　27位
テストケース絞り込みにおける留意事項（テストケースを絞り過ぎて障害に至った事例紹介）	3件	絞る　＆　テスト ※	複合検索不能	適合　9件 検索　144件 適合最低　33位
開発環境と本番環境の違いによる障害事例紹介	3件	開発環境と本番環境の違い	複合検索不能	適合　8件 検索　100件 適合最低　28位

※　ベクトル空間モデルを利用した検索（＝類似文書の検索）では、アンド条件の検索ができないので、2段階検索の機能を追加した



有用性①　類義語検索により想定外の障害リスクを検知（類義語検索の効果1）

文字列検索とAI検索エンジンの件数差異分析

検索語	文字列検索	AI検索（類義語20語）	開発ノウハウ（障害事例）
文字コード	12件	適合　15件 検索　30件 最低　27位	①SJISの亜種による障害事例 ②SJISのバージョン違いによる障害事例 ③外字問題（端末に外字未登録） ④外字問題（他システムの外字が未登録）　4件

文字列検索との差異3件（＝”文字コード”を含まない障害）の内容

順位	障害内容	開発ノウハウの記載	検索でヒットした類義語
9	変換不能時の代替文字を指定する箇所が2箇所存在していたが、それに気付かず1箇所のみと思い込み作業をしていた。	有　想定内障害で、変換不能時の注意事項を記載	変換
13	半角カナ文字を含むデータ項目について、仕様上文字数×3倍のバイト数を指定すべきであったが、文字数と同じ長さで定義してしまっていた。（UTF－8化の対応漏れ）	無　想定内障害であるが、基本事項ため記載せず	コード変換
17	EBCDICコード領域にSJISコードのSPACE(0X20)が設定されていたことから、変換エラー発生	無　想定外障害のため、開発ノウハウに記載なし	SJIS コード変換

開発ノウハウ作成時には気付かなかった障害リスクを、AI検索エンジンで検索できた！

有用性② 検索語の特定が難しい障害の検索に効果大（類義語検索の効果2）

「テストケースを絞り過ぎて障害に至った事例」を紹介する開発ノウハウ作成時の障害調査と、AI検索の比較

検索語	開発ノウハウで紹介した障害事例
適当な検索語でひたすら検索	①共通モジュールテストのテスト漏れ ②画面と帳票の片方だけテスト ③分析サーバの性能テスト漏れ

かなり苦労して障害を抽出し、開発ノウハウを作成

VS

検索語	類義語	AI検索
絞る ＆ テスト	20語	適合 9件 適合最低 33位 検索 157件
同上	無し	適合 1件 適合最低 1位 検索 4件

AI検索エンジンによる検索結果の紹介

順位	検索結果
1	担当者は特定月に設定される特定項目の確認に気を取られて、特定項目以外の項目に対する確認への意識が希薄であった。
2	発行通数1通をベースにテストしたため、改ページ条件の誤りを検知できなかった。
7	○○が実施される月かそれ以外の月の双方をテストすべきところを、いずれかの月のみをテストすれば十分と判断してしまった。
14	現行踏襲であり、各項目の設定内容やデータシーケンスに変更はないため確認不要と判断してしまった。
33	○○機能は共通化された処理であることから、いずれか代表機能でテストすればよいと判断し、バリエーションを絞ったテストに留めたためにテストで検知せず、、

「絞る」「テスト」だけで多くの参考になる障害を検索できる

開発ノウハウの障害事例①と同じ

3. 検証結果　②利便性 1/3

利便性①　ネット検索のように、複合検索ができ、検索語と適合した文書が上位に検索された

検索結果例の紹介　：　性能障害の検索結果

類義語検索なしの検索結果

検索語（入力内容）	実際の検索語（形態素解析後）
CPU使用率　メモリ　不足　　レスポンス遅延	CPU　　使用　　率　メモリ　不足　　レスポンス　　遅延

順位	検索結果
1	○○のレスポンス遅延発生。全体のCPU使用率を高め、90%～100%の高騰状態となったもの。
2	オンライン専用機がCPU使用率90%超。発生条件が特定されていない○○製品の不具合。
3	保証料率=0.001%の条件の他に、保証料率=NULL値の条件が存在。
4	ORACLEのメモリ使用率が上昇し、ページアウトが多発し、サーバが高負荷状態になり、、、
29	○○エラーに伴う電文処理遅延。メモリ不整合により、プロセスがアップダウンを繰り返す。
30	セッションが上限に達した。CPU、メモリ、ハードディスクのリソースが逼迫していた形跡もなく、、

上位に検索語に適合した障害が出力されている！

検索語の”率“に反応している

順位が低くなると性能障害以外の障害が出力

適合文書数　　上位　7/10文書　上位　11/20文書　　上位　15/30文書

上位文書に適合文書が多い

類義語20語有の結果
上位　8/10文書
上位　12/20文書
上位　18/30文書

検索語が多いと、類義語効果は低下

利便性② 適合しない文書を除く工夫1

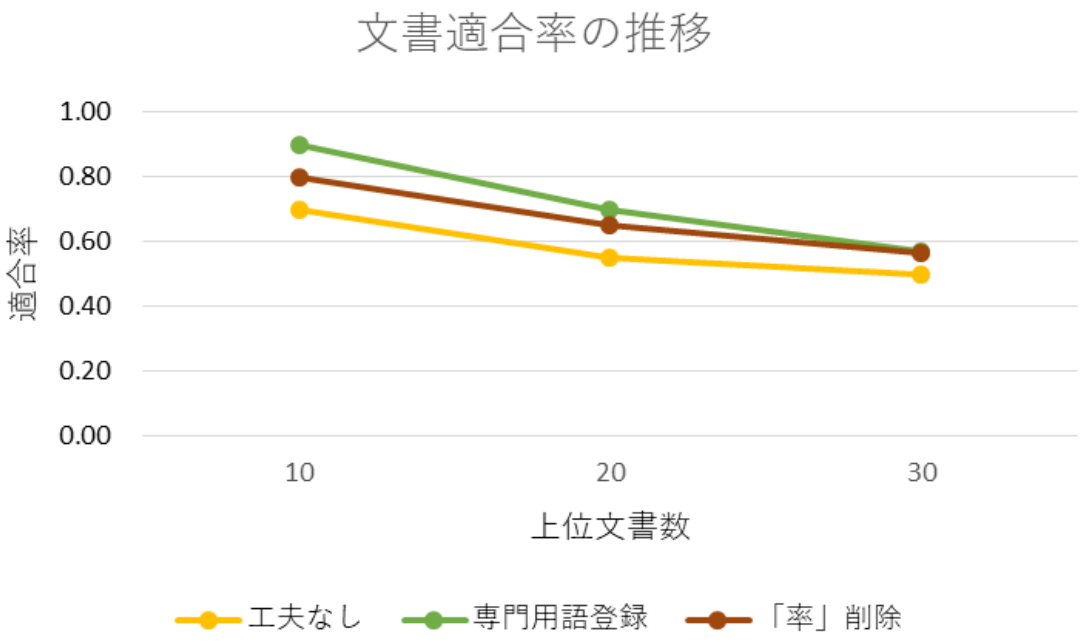
検索語（入力内容）	実際の検索語（形態素解析後）
CPU使用率　メモリ　不足　レスポンス遅延	CPU　使用　率　メモリ　不足　レスポンス　遅延

順位	検索結果
3	保証料率=0.001%の条件の他に、保証料率=NULL値の条件が存在。

性能障害に適合しない障害

方法1　専門用語「使用率」の登録

方法2　検索語から「率」を除外する



- ✓ **専門用語を登録する方が効果が高い**
  ただし、専門用語の登録は諸刃の剣
  「使用」で検索した場合、「使用率」は検索されない
- ✓ **検索用語の除外でも効果はある**
  検索語（類義語を含む）の取捨選択機能は効果的

利便性③ 適合しない文書を除く工夫2

類義語検索の精度を上げて適合率を高める。そのためにWord2Vecを追加学習する。

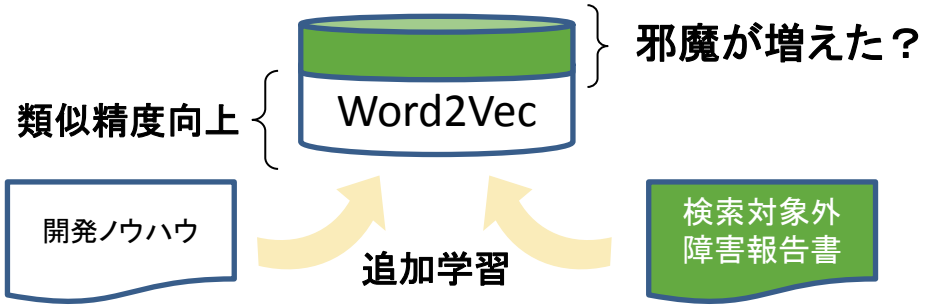
	内容	ポイント
方法1	開発ノウハウを追加学習	開発ノウハウは障害報告書から作成しているので、検索対象文書と同じ単語集合体で学習
方法2	検索対象外の障害報告書を追加学習	検索対象文書と異なる文書で学習

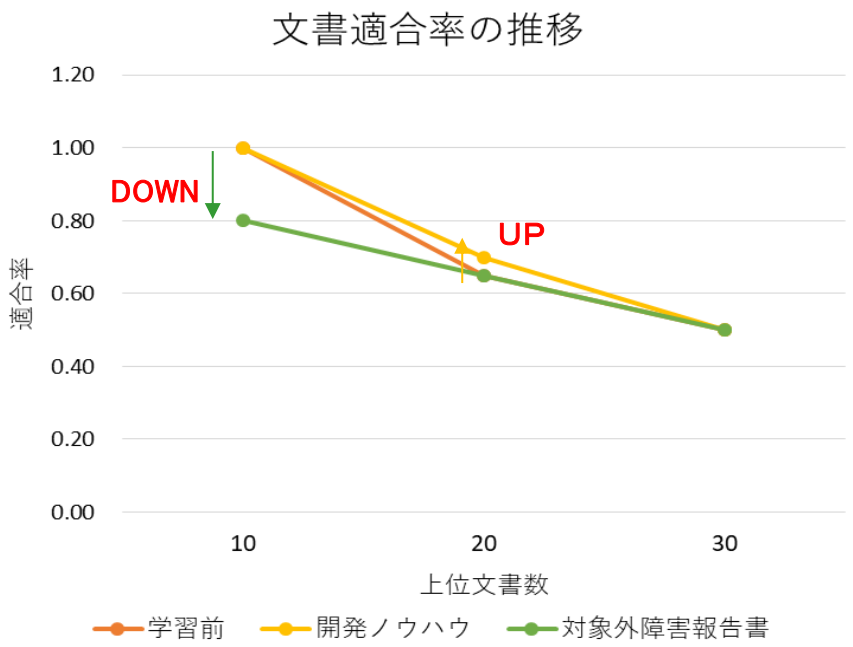
✓ **検索対象外の文書から学習すると、返って精度が低下する可能性がある**

新たに類義語が増えても、

- **検索でヒットしない無駄な類義語が上位に入り込む（邪魔）**
- **検索テーマと関係ない障害とヒットする（足を引っ張る）**

✓ **検索対象文書と同じ単語集合体で追加学習できる方法を考えた方が良い**





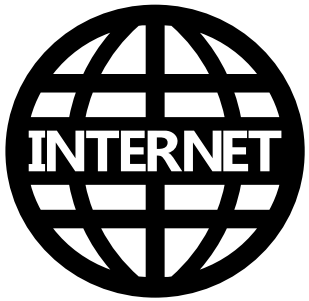
文字コード（類義語20語）の検索結果

4. まとめ　①問題解決が期待できるか？

問題点	効果（検証結果）	問題解決の貢献内容
提供側の問題 可視化した**開発ノウハウは障害報告書の8%程度**に留まっている	**有用性** ✓ 開発ノウハウの水準以上の障害件数とバリエーションを検索できた ✓ 開発ノウハウ可視化のために苦労した障害調査が、容易に検索できた（検索語の特定が難しい障害調査に効果大）	✓ 開発ノウハウにおける「障害リスクの注意喚起」という役割においては、AI検索エンジンによる障害報告書の検索が習慣化されれば、ある程度代替可能と考える ✓ 開発ノウハウ可視化のための障害調査が容易になるため、開発ノウハウ可視化が加速する
利用者側の問題 **開発ノウハウへのアクセスが少なく活用度が低い**	**利便性** ✓ ネット検索のように、検索語と適合した文書が上位に検索された 紹介した4件の検索例の上位10件の適合率は平均68%	開発ノウハウと障害報告書を一緒にDBに入れ、障害報告書と共に開発ノウハウも検索されるようにすれば、開発ノウハウへのアクセスが増えると期待できる

4. まとめ　②展望と課題

～こんな風に活用されることを期待している～



文字列を数値に変換したい。ネットで検索しよう

開発ノウハウや障害報告書で、障害リスクも確認しておこう！
「文字列、数値変換」

文字列変換で端数誤差が生じる障害報告書、開発ノウハウが上位でヒット！

同原因の再発防止策を検索して、再発防止策の妥当性を確認しよう！

設計作業中
再発防止策検討

開発ノウハウ
障害報告書

AI検索エンジン

社内の他文書

色々な文書で利用可能
（例）顧客満足度評価、事務手続等の規定・手続

AI検索エンジン導入後に必要と考えていること

- ✓ 活用事例（検索例）の紹介
- ✓ 開発ノウハウ、障害リスクの理解度確認テストの実施
  （テストを通じて検索を体験）