IT システム開発における類似プロジェクト検索技術の開発

山本 智基[1]　伏田 享平[1]　滝本 雅之[1]

概要：IT システム開発プロジェクトにおいて類似するプロジェクトの情報を参照する機会は多い．開発，管理に従事する要員は，これらの情報を用いてプロジェクトの計画作成や IT システムの設計，実装を行う．一方で，プロジェクトの類似性を判断する観点は明らかになっていない．そのため類似プロジェクトの選定作業は属人的になっている．本稿では IT システム開発において，類似プロジェクトの選定，参照を支援する仕組みを提案する．まず，どのような観点に着目してプロジェクトの類似性を判断しているかを明らかにするため，実務者を対象にアンケート調査を実施した．調査の結果，利用シーンにより重要視される観点に一定の傾向があることを確認した．次に調査結果をもとに，プロジェクトの類似性を定量的に表す指標を定義した．この指標は類似プロジェクトを利用するシーンに応じて，類似性を判断する観点の重要度を考慮できる．最後に，提案する指標を用いて類似プロジェクトを検索する Web システムを構築した．この Web システムを用いることで，熟練者でなくても利用シーンに応じて容易に類似プロジェクトを検索，参照することが可能となる．

1. はじめに

IT システム開発プロジェクトにおいて，類似するプロジェクトの情報を参照し，プロジェクトの計画，管理，実行する機会は多くある [7]．例えばプロジェクトの計画時には，同一の規模やシステムアーキテクチャによる過去プロジェクトの実績値をもとに工数見積もりを実施する場合がある．また，類似プロジェクトでのプロジェクト実行上の留意点をあらかじめ把握しておくことで，円滑なプロジェクト運営が期待できる．

一方でプロジェクトの類似性を判断する観点は明らかになっていない．例えば，開発計画時に自プロジェクトの生産性を検証したい場合には，同規模のプロジェクトを類似プロジェクトとみなす場合がある．また，新規に採用する開発ツールを用いた開発方法を参照したい場合には，同様のツールを用いたプロジェクトを類似プロジェクトとみなす．このように類似プロジェクトの選定作業はプロジェクトを参照する目的に応じた観点を定める必要があり，その作業自体は属人的となっている．

類似性の判定にあたっては熟練者によって判断することが望ましい．例えば，大規模な開発では一般的に生産性は低下する．一方で，対象としているシステムの利用用途のみに着目すると，規模に見合わない高い生産性を見込むおそれがある．このように利用シーンやプロジェクトの状況により，類似性を判断する観点が異なるため，熟練者による適切な判断が必要となる．

本稿では IT システムプロジェクトにおける類似プロジェクトの選定，参照を支援する仕組みを提案する．まず，類似プロジェクトを参照する利用シーンを整理し，それぞれの利用シーンでどのような観点でプロジェクトの類似性を判断しているか調査した．次に，調査結果をもとにプロジェクト間の類似性を表す定量的指標を定義した．さらにこの指標をもとに，類似プロジェクトを検索するシステムを開発した．このシステムは Web システムの形で実装し，NTT データ社内に展開している．

本研究における貢献は下記の 3 点に集約される．

- 利用シーン別に，実務者がプロジェクトの類似性を判断する際に重要視する観点を明らかにした．
- 実務者が類似性を判断する際の観点をもとに，プロジェクトの類似性を定量的に表した指標を定義した．
- 規模や利用ツールといったプロジェクトのプロフィールをもとに，類似するプロジェクトの情報を検索するシステムを開発した．

上記の貢献により，熟練者でなくても利用シーンに応じて容易に類似プロジェクトを参照することが可能となる．

以降，2 節では類似性を判断する観点について調査した結果を述べる．3 節では 2 節で示した調査結果をもとに定義したプロジェクト間の類似性を表す指標について述べる．4 節では 3 節で示した指標をもとに類似プロジェクトを検索するシステムの開発と評価について述べる．5 節で関連研究を示し，6 節でまとめを述べる．

[1] 株式会社 NTT データ

表 1 類似プロジェクトの情報を参照する利用シーン

	タイミング	参照する目的	参照する情報
U1	提案時（受注前）	提案書の作成	提案資料，プロジェクト概要
U2	プロジェクト計画	見積もり作成	生産性の実績データ
U3	プロジェクト計画	プロジェクトが見積もった結果の妥当性検証	生産性の実績データ
U4	プロジェクト計画	品質の基準値を設定する際の目安の把握	品質の実績データ
U5	開発中	システムの実現方式や設計方式の検討	設計書などの開発成果物
U6	開発中	品質の分析	品質の実績データ
U7	プロジェクト振り返り	生産性向上の糸口を探す	開発実績データ

表 2 プロジェクト属性

プロジェクト属性	概要	属性値の例
顧客業種	プロジェクトの顧客が所属している業種	金融業，流通・サービス業
業務内容	対象とする業務内容	勘定系，情報系
開発区分	開発したシステムの開発種別	新規開発，更改開発，機能拡充開発
案件区分	プロジェクトの開始形態	受託型，自社企画型
開発規模	開発したシステムの規模	100K LoC
生産性	開発に関する生産性	規模単位/人月，円/規模単位
工数	開発にかかった工数	50 人月
開発期間	開発にかかった期間	10 ヶ月
開発体制	プロジェクトの平均要員数	50 人
開発工程	プロジェクトで実施した開発工程	要件定義〜試験，外部設計〜試験
開発形態	システムの開発形態	フルスクラッチ，パッケージ利用
使用言語	利用した開発言語	Java，Python
手順・ツール	用いた開発手順，開発ツール	SAP，Salesforce

2. プロジェクトの類似性を判断する観点

本節では実務者が類似プロジェクトの情報を参照する利用シーンと，その際にプロジェクトの類似性を判断する観点であるプロジェクト属性について示す．その上で，実務者が類似プロジェクトを参照する際に特に着目しているプロジェクト属性を明らかにするため実施したアンケート調査について述べる．

2.1 類似プロジェクトの利用シーンとその判断基準

表 1 に本研究で想定している類似プロジェクトを参照する利用シーンを示す．プロジェクトが発足する前の提案活動時からプロジェクト計画，実行中とあらゆるタイミングで類似プロジェクトの情報は参照される．また，類似プロジェクトを利用するタイミングに応じて，その参照する目的や情報も異なってくる．類似プロジェクトを参照する際には，利用目的と自身のプロジェクトの状況をふまえて，自身のプロジェクトと類似したプロジェクトを選定する．

過去の類似プロジェクトとの類似性を判断する観点として，プロジェクト属性が挙げられる．**表 2** に本研究で想定しているプロジェクト属性を示す．プロジェクト属性とは，多様なプロジェクトの特性を表現するために定義されたデータ項目である．本研究において，プロジェクト属性の選定は，実務者および社内の有識者の意見を参考に決定した．

2.2 調査内容

実務者が実際に類似プロジェクトを利用するユースケースと，類似性を判断する観点を明らかにするため，アンケート調査を行った．アンケートでは下記の設問を設けた．

類似プロジェクトの利用経験 類似プロジェクトの情報を利用したことがあるか．経験がない場合，今後参照したいか．

類似プロジェクトの利用シーン 実際に類似プロジェクトの情報を参照した経験がある場合，表 1 に示した利用シーンのうちどの利用シーンで利用したか．

類似プロジェクトの判断観点 類似プロジェクトの情報を参照した際，プロジェクトの類似性を判断する上で重要視する観点は表 2 のプロジェクト属性の中で何か．

いずれの設問も複数回答可能とした．上記に加えて，類似プロジェクトを参照する際に必要な情報，要件について自由記述式でコメントを収集した．調査は NTT データおよびそのグループ会社に所属する社員を対象に，社内のオンラインアンケートシステムを用いて 2017 年 6 月に実施した．

2.3 調査結果

回答者数は 115 人であった．**表 3** に，アンケート回答者のプロジェクトでの役割ごとに利用経験の有無を示す．「経験有」はいずれかの利用シーンで参照した経験があると回答した回答者の数を指している．「意欲有」はいずれの利用

シーンでも参照した経験はないが，今後いずれかのシーンで参照したいと回答した回答者の数を指している．「意欲不明」はいずれの利用シーンでも参照した経験がなく，今後参照したいかわからないと回答した回答者数を指している．回答者の多くはプロジェクトマネージャ（PM）や開発リーダといった，開発プロジェクトを主導する役割である．これらのメンバは主にプロジェクト管理を実施する．その他には，庶務系のスタッフや営業職が含まれていた．

役割別に見ると，最も利用経験がある回答者の割合が高い役割は PM である．これは，プロジェクトを管理する立場であり，自身のプロジェクトを管理する上で類似プロジェクトの情報を活用する場面が多いためと考えられる．全ての利用シーンにおいて，類似プロジェクトの情報を参照した経験がなく，今後も参照しないと回答した回答者は 0 名であった．この結果より，類似プロジェクトの情報を利用するニーズが高いことが確認できる．

過去に類似プロジェクトの情報を参照した回答者が，その情報を利用した利用シーンを**表 4** に示す．利用シーンによって類似プロジェクトの情報が利用される割合（利用率）は異なる傾向がある．最も利用率が高い利用シーンである U2（プロジェクト計画時の見積もり作成）は，最も利用率が低い利用シーンである U7（プロジェクト振り返り時の参考情報参照）よりも 6.5 倍も利用率が高い．利用タイミングに着目すると，プロジェクトの計画以前の方が利用率が高い．理由として，プロジェクトの序盤ほど不確定要素が多く，それらを検討する際に類似プロジェクトの情報を参照することが多いためと考えられる．

表 5 に各利用シーンにおいて，類似プロジェクトと判断するために重要視したプロジェクト属性を示す．「顧客・顧客業種」が重要視される割合に着目すると，U1（提案時の提案書作成）では 65%であるのに対し U4（プロジェクト計画時の品質基準値の把握）では 21%であった．一方で，「業務内容」や「開発形態」は全ての利用シーンで重要視されやすいプロジェクト属性であることがわかる．これらのことから，利用シーンによって重要視されやすい観点が異なると考えられる．

2.4 考察

調査結果から，類似プロジェクトの情報の利用シーンによって，プロジェクトの類似性を考慮する上で重要視するプロジェクト属性が異なる傾向があると考えられる．加えて，複数の回答者から「利用したい情報が同じであっても，自身が置かれた立場，プロジェクト状態によって，重要視するプロジェクト属性は異なる」といった回答があった．類似プロジェクトの情報を参照する際の要件としては，「より多くの類似プロジェクトの情報を参考にしたい」，「プロジェクト属性を入力したら、入力条件に近いプロジェクトでの生産性や品質実績データの統計値を取得できるようにしてほしい」といった意見が存在した．

表 3 回答者のプロジェクトでの役割

役割	人数	経験有	意欲有	意欲無	意欲不明
PMO	7	2	4	0	1
PM	30	27	2	0	1
開発リーダ	46	30	15	0	1
開発メンバ	21	13	7	0	1
その他	11	9	1	0	1

表 4 類似プロジェクトの利用シーン

役割	U1	U2	U3	U4	U5	U6	U7
PMO	0	2	2	2	0	1	0
PM	20	26	17	16	15	8	5
開発リーダ	20	21	21	12	17	12	2
開発メンバ	5	11	10	3	6	2	2
その他	7	5	4	5	3	2	1
合計	52	65	54	38	41	25	10

表 5 重要視するプロジェクト属性

プロジェクト属性	U1	U2	U3	U4	U5	U6	U7
顧客・顧客業種	34	26	22	8	13	5	4
業務内容	42	47	41	17	25	14	8
開発区分	36	44	41	27	21	18	4
案件区分	18	18	18	14	9	9	1
開発規模	29	45	35	28	21	16	5
生産性	7	25	21	14	10	10	2
工数	9	22	21	16	6	7	4
開発期間	21	25	22	14	13	8	3
開発体制	10	17	17	12	11	10	2
開発プロセス	23	29	28	20	20	12	5
開発形態	30	42	41	27	25	17	8
使用言語	22	34	27	22	21	13	6
手順・ツール	18	29	26	18	19	12	3

本調査より，プロジェクトの類似性を示す指標定義，並びに，類似プロジェクト検索システム開発の際は下記の 2 点を考慮すべきであると示唆される．

- 利用シーンによって，重要視されるプロジェクト属性の度合いが異なる傾向があること
- 人や状況，役割によって，重要視するプロジェクト属性が異なる傾向があること

3. プロジェクト類似度

前節での調査結果より，プロジェクトの類似性判定にあたっての観点は一定の傾向が見られた．これより上記の結果をふまえてプロジェクトの類似性を定量的に表すことで，類似プロジェクトを客観的に判定することができると考えられる．本節ではプロジェクト間の類似性を定量的に表す指標（以下，「**プロジェクト類似度**」と呼ぶ）の定義について述べる．

3.1 基本アイデア

前節で示した調査結果より，プロジェクトの類似性判定にあたっては複数の観点，すなわちプロジェクト属性をもとに判断していることがわかる．また，類似プロジェクトの参照目的に応じて，どのプロジェクト属性を重視するかは異なる．よって，類似プロジェクトを参照する目的に応じて重視するプロジェクト属性を定め，それに応じた類似度を算出できることが望ましい．

そこで本研究ではプロジェクトを d 個の要素より構成される d 次元ベクトルとみなす．ここで各要素は，当該プロジェクトの工数や開発区分といったプロジェクト属性とみなせる．その上で，プロジェクト間の距離を類似度とみなすこととした．上記に基づく本研究におけるプロジェクト類似度の基本アイデアを下記に示す．

プロジェクト類似度の基本アイデア

プロジェクト属性が $\mathbf{c}$ であるプロジェクトを P_S としたとき，プロジェクト群 $\mathcal{P} := \{P_1, P_2, \ldots, P_N\}$ を d 次元実ベクトル空間の点に移すような写像 $f_c : \mathcal{P} \to \mathbf{R}^d$ を作る．この写像は，以下の条件を満たす．

- P_S と完全に一致する，すなわちプロジェクト属性が $\mathbf{c}$ であるプロジェクト P_i に対し，$f_c(P_i) = \emptyset$ (原点) を満たす．
- P_S と類似するプロジェクト P_i に対し，$f_c(P_i)$ は $\emptyset$ に “近い” 点となる．

上記の条件のもとで，プロジェクト P_i のプロジェクト類似度を，$\emptyset$ と $f_c(P_i)$ との距離で定義する．

上記の基本アイデアを実現するためには，f_c を定義する必要がある．プロジェクト属性は，案件区分，言語などのカテゴリデータと，規模，工数などの定量的データに分けられる．カテゴリデータは，プロジェクト属性によって取りうる要素数が異なる．例えば，「案件区分」の場合であれば，その取りうる要素は { 受注型，企画型，その他 } の 3 種類だが，「言語」であれば 15 種類存在する．そのため，要素数が多いプロジェクト属性ほど検索に影響を与えやすくなる．定量的データは，プロジェクト属性によって取りうる値の範囲が異なる．そのため，値が取りうる範囲が広いプロジェクト属性ほど検索に影響を与えやすくなる．

これらの点に考慮し，プロジェクト属性ごとに 1 次元化を行い，値域が [0,1] である類似度関数を定義する．その類似度関数を用いた f_c を次のように定義した．

表 6 プロジェクト一覧の例

	案件区分	言語
P_1	受注型	Java, C
P_2	企画型	C++

表 7 {0,1} ベクトル空間化変換後の例

	案件区分			言語		
	受注型	企画型	その他	Java	C	C++
P_1	1	0	0	1	1	0
P_2	0	1	0	0	0	1

写像 f_c の定義

検索条件として指定されたプロジェクト属性を $\mathbf{c} = (c_1, c_2, \ldots, c_d)$，あるプロジェクト P_i のプロジェクト属性を $\mathbf{e} = (e_1, e_2, \ldots, e_d)$ とする．このとき，写像 $f_c(P_i)$ を

$$f_c(P_i) := (Sim(c_1, e_1), \ldots, Sim(c_d, e_d)) \qquad (1)$$

と定義する．ここで Sim は，プロジェクト属性がカテゴリデータならば sim_{cat}，定量的データならば sim_{qua} とする．

以降，カテゴリデータに対する類似度関数 sim_{cat}，定量的データに対する類似度関数 sim_{qua} を定義する．その上で，類似度を算出するための距離関数を示すことで，プロジェクト類似度を定義する．

3.2 カテゴリデータに対する類似度

一般にカテゴリデータを数値情報に変換する手法として，{0,1} ベクトル空間に書き換えることが多い．ここでは，まず各プロジェクトのカテゴリデータ（**表 6**）を，{0,1} ベクトル空間（**表 7**）へ変換する．その後，プロジェクト属性の要素数の違いを考慮するために，各プロジェクト属性を 1 次元化する．カテゴリデータに対する類似度関数 sim_{cat} を以下のように定義する．

カテゴリデータに対する類似度関数の定義

あるプロジェクト属性 e について，プロジェクト a, b の値を e_a, e_b とする．また，その {0,1} ベクトルを，$\mathbf{e}_{bin_a}, \mathbf{e}_{bin_b}$ とする．このとき，$sim_{cat}(e_a, e_b)$ を

$$sim_{cat}(e_a, e_b) = 1 - \frac{\mathbf{e}_{bin_a} \cdot \mathbf{e}_{bin_b}}{|\mathbf{e}_{bin_a}||\mathbf{e}_{bin_b}|} \qquad (2)$$

と定義する．

この定義より，カテゴリデータが検索条件と完全一致するプロジェクトの場合は 0 となる．例えば，表 7 について e を案件区分としたとき，$sim_{cat}(e_{P_1}, e_{P_2}) = 1$ となる．

カテゴリデータに対する類似度は，与えられた検索条件によらず事前に構成しておくことが可能である．よって 4 節で述べる類似プロジェクト検索システムにおいては，カ

表 8　案件区分の類似度行列の例

	受注型	企画型	その他
P_1	0	1	1
P_2	1	0	1

テゴリデータごとにあらかじめ類似度を算出し保持している．算出時は，プロジェクト属性ごとに検索条件が取りうる要素パターンを生成して，プロジェクト属性の類似度行列を生成する．上記をふまえ，カテゴリデータに対する類似度行列を以下の通り定義する．

カテゴリデータに対する類似度行列の定義

あるプロジェクト属性 e の要素に対する類似度行列を $T = (t_{ik})$ とする．また，e_{P_i} を e に関する P_i の要素，u_k を e が取りうる要素パターンとする．このとき，

$$t_{ik} = 1 - sim_{cat}(e_{P_i}, u_k) \tag{3}$$

である．

上記の定義において，要素パターン u_k は e の要素の数だけ生成される．例えば，案件区分が取りうる要素が { 受注型, 企画型, その他 } である場合，k = { 受注型, 企画型，その他 } となる．要素パターンを {0,1} ベクトル化したものをそれぞれ $\mathbf{u}_{order}$，$\mathbf{u}_{plan}$，$\mathbf{u}_{other}$ とすると，$\mathbf{u}_{order} = (1,0,0)$，$\mathbf{u}_{plan} = (0,1,0)$，$\mathbf{u}_{other} = (0,0,1)$ となる．**表 8** に式 (3) に基づき表 7 の案件区分について類似度行列を生成した結果例を示す．

3.3　定量的データに対する類似度

定量的データはプロジェクト属性によって取りうる値域が異なる．ここでは，カテゴリデータの類似度が取りうる値域 [0,1] と同様な範囲へ定量的データの値を変換する．この変換には下記のように定義した累積確率を利用する．

定量的データに対する類似度の定義

検索条件として指定された定量的データの値を m_S とする．検索対象の各プロジェクト $P_1, \ldots, P_N$ において，検索条件として指定された定量的データの値を $m_1, \ldots, m_N$ とする．また，$m_1, \ldots, m_N$ の任意の値を m_{target} とする．このとき，類似度 $sim_{qua}(m_S, m_{target})$ を以下のように定義する．

(1) $|m_1 - m_S|, |m_2 - m_S|, \ldots, |m_N - m_S|$ で確率密度関数を作る．

(2) (1) で作った確率密度関数の累積分布関数 $p(m_{target})$ を定量的データに対する類似度関数 $sim_{qua}(m_S, m_{target})$ と定義する．

この定義より，検索条件と完全一致するプロジェクトの場合はその類似度は 0 を取り，検索条件と値の差が大きくなるほど，類似度は大きくなる．

3.4　Mahalanobis 距離によるプロジェクト類似度

3.1 節で述べたように，本研究では $\emptyset$ と $f_c(P_i)$ との距離をプロジェクト類似度として扱う．そして，距離の値が小さいほど類似プロジェクトであるとみなす．一般的に，プロジェクトの実績データには分布の偏りが存在する [5]．Euclid 距離では多くのプロジェクトが集中するデータ群とそうでないデータ群との差異が考慮されず，類似度として扱うには不適と考えた．そこで分布の偏りを考慮した Mahalanobis 距離を本研究では採用する．

個体 i の Mahalanobis 距離 MD_i は，$\mathbf{p}_i$ を個体 p_i の特徴ベクトル，$\bar{\mathbf{p}}$ を個体群の平均ベクトル，$\mathbf{\Sigma}$ を共分散行列とすると，下記のように定義される。

$$MD_i = \sqrt{(\mathbf{p}_i - \bar{\mathbf{p}})\mathbf{\Sigma}^{-1}(\mathbf{p}_i - \bar{\mathbf{p}})^{\mathrm{T}}} \tag{4}$$

なお，T は転置ベクトルを意味する．

ここでは検索条件からプロジェクト類似度を計算する概要を説明する．まず，検索条件が与えられたとき，写像 f_c を用いて N 個のプロジェクト群を変換し，類似度算出用の行列 $S(N \times d)$ を生成する．類似プロジェクトの参照目的に応じたプロジェクト属性の重要度合いを考慮するため，$S' = SW$ とし，S' に対して共分散行列 $\mathbf{\Sigma}$ を算出する．なお，$W(d \times d)$ はプロジェクト属性の重み行列であり，プロジェクト属性 e の重み w_e を対角要素とする対角行列である．S' に対して算出した共分散行列を使用することで、重要視するプロジェクト属性 e のプロジェクト類似度への寄与を強くすることができる．

4 節で述べる類似プロジェクト検索システムにおいては，w_e として 2 節の調査結果，および有識者へのヒアリング結果に基づき設定した値を用いている．また，カテゴリデータについて属性ごとに生成した類似度行列から，検索条件と合致する列ベクトルを抽出して S を生成する．その際，検索条件に複数の属性値が設定されている場合は，抽出した列ベクトルの平均値を用いる．

4.　類似プロジェクト検索機能の開発

本節では 3 節で定義したプロジェクト類似度を用いて類似プロジェクトを検索するシステムの構築について述べる．まず類似プロジェクト検索機能のユースケースと機能要件について述べる．その上で，今回開発した検索機能とその評価について示す．

4.1　機能要件の整理

検索機能の仕様を詳細化するため，前述したアンケート調査結果をもとに NTT データに所属するプロジェクトマネージャ 6 名に対してヒアリングを行った．いずれの回答者もプロジェクト管理に関わる社内資格を保有している．

ヒアリングにあたって，機能のユースケースとしてアンケート調査結果から最も多くの人が類似プロジェクトの情

報が利用した経験のある U2（プロジェクト計画時の見積もり作成）を想定した．ヒアリングでは，U2 に加え U3（生産性の検証作業）もできるとよいという意見が多数出た．

ヒアリングの結果から，本機能のユースケースと満たすべき機能要件を下記のように定義した．

ユースケース　見積もり作成時，設定した生産性計画値の検証のために，類似プロジェクトの情報を参考にする．検証時は，まず設定した計画値が類似プロジェクトの実績値と乖離していないか確認する．その後，類似プロジェクトの詳細情報を確認する．

要件 1　1 回の検索行動で多くの類似プロジェクトの情報にアクセスできる

要件 2　利用者が重要視するプロジェクト属性を任意に設定可能で，検索結果に反映される

要件 3　類似プロジェクトの実績値の箱ひげ図が表示され，自身が設定した計画値との乖離具合を確認できる

要件 4　検証するために類似プロジェクトの詳細情報を閲覧することができる

以降，上記のユースケースと要件のもと開発した類似プロジェクト検索機能について詳細を紹介する．

4.2　類似プロジェクト検索システムの構築

NTT データでは，過去のプロジェクトの生産性や品質の実績データを参照できる Web システムを構築し，社内で運用している [11]．このシステムでは，ユーザが規模や採用しているプログラミング言語といったプロジェクト属性を指定すると，指定した条件に合致するデータを検索し，散布図の形式で表示する．本研究ではこのシステムを機能拡張することで類似プロジェクト検索機能を開発した．

類似プロジェクト検索機能では，ユーザが所属するプロジェクトのプロジェクト属性と，ユーザ自身が重要視するプロジェクト属性を指定し，検索を実行する．3 節で示し

図 1　入力画面

図 2　検索結果画面

たプロジェクト類似度を用いて，類似したプロジェクトを特定する（**要件 1**）．検索結果画面には類似プロジェクトの情報が表示される．

検索対象データ

NTT データではプロジェクト完了時に，プロジェクト属性や生産性，品質に関する実績データを収集している [12]．本検索機能ではこれらの実績データを検索対象としている．検索対象のプロジェクト件数は，約 4,000 件である．なお検索機能で用いるプロジェクト属性は，データ欠損を考慮し，現時点では「顧客区分」「顧客業種」「案件区分」「開発区分」「基盤区分」「サービス種別」「使用言語」「開発規模」の 8 属性とした．

検索条件入力画面

図 1 に検索条件入力画面を示す．ここでは，ユーザは自身が所属するプロジェクトの属性情報を選択する．また，重要視するプロジェクト属性が存在する場合，そのプロジェクト属性の重要視することを指す「重要視フラグ」を選択する（**要件 2**）．重要視フラグを設定すると，3.4 節におけるプロジェクト属性の重みが定数倍される．加えて生産性 [KS/人月] の計画値が存在する場合はその値を入力する．ユーザが入力するプロジェクト属性は全て必須入力としている．

検索結果画面

ユーザの検索条件をもとに，類似プロジェクトの情報を表示する．検索結果画面を**図 2** に示す．検索結果画面では，散布図，箱ひげ図，統計量，類似プロジェクトの一覧表が表示される．

箱ひげ図には類似プロジェクトの実績値に自プロジェクトの計画値が重ねて表示される．これによりユーザは計画値が類似プロジェクトの実績値の分布から乖離していないかを確認することができる（**要件 3**）．

類似プロジェクト一覧表には，プロジェクト類似度順に類似プロジェクトとそのプロファイル情報が一覧表示される．加えて，プロジェクトを指定するとプロジェクトの詳細情報を閲覧することができる．これによりユーザは自身のプロジェクトと各類似プロジェクトを比較しながら，計画している生産性の値が妥当であるか検証することが可能となる（**要件 4**）．

類似判定処理

箱ひげ図（**要件 3**）を作成するにあたって，類似プロジェクトの統計量を算出する必要がある．統計量を算出するために，検索条件に対するプロジェクト類似度が算出されたプロジェクト群から，類似プロジェクトの件数を決定する必要がある．しかし，算出したプロジェクト類似度の分布は検索条件によって変化する．そこで，3.4 節で算出されたプロジェクト類似度の分布から，類似プロジェクトの判定を行う．まず，プロジェクト類似度順に 1 位，2 位，. . . , N 位のランク付けを行う．その後，プロジェクト類似度が

急激に変化するランク θ（**図 3** 中の縦線）を検知した後，θ 位までのプロジェクトを類似プロジェクトとして判定する．

θ の検知方法は経験的に策定した．まず，プロジェクト類似度の対数に対して回帰を行う．次に，回帰曲線とプロジェクト類似度との残差に対してスプライン回帰を適用し，図 3 中の赤線に示すプロジェクト類似度の傾向の変化点を求める．一方で，必ずしも θ を抽出できる保証がないため，抽出できなかった場合は，θ=100 としている．

4.3 機能評価

開発した検索機能について，想定ユースケースにおける検索結果の傾向を分析，評価する．ここでは、過去のプロジェクトの情報を用いて検索を行い，検索結果件数とその検索結果から得られる生産性の 25 パーセンタイルと 75 パーセンタイルに当該プロジェクトの生産性の実績値が入っているかという観点で机上検証を行うことにした．

機能要件の整理を行う際に実施したヒアリングにおいて，上記のような観点で自プロジェクトの計画値と実績値の比較を行い，生産性の妥当性を検証しているとの回答が得られた．よってこのような作業の対象となる類似プロジェクトの件数がどのように変化するか確認するため，上記のような観点で機能評価を行った．あわせて本検索機能の使用感について利用者へアンケート調査を行った．

机上検証

机上検証における提案手法の比較対象として，文献 [11] のシステムで用いられている検索方式を採用した．文献 [11] のシステムでは，検索時には複数のプロジェクト属性を指定でき，全てのプロジェクト属性に一致するプロジェクトが検索結果として表示される．（以降，このような検索方式を「AND 検索」と呼ぶ．）文献 [11] のシステムの運用を通して，AND 検索では指定したプロジェクト属性の種類が増えると検索結果として表示されるプロジェクトの数が顕著に減少する傾向にあることがわかっている．

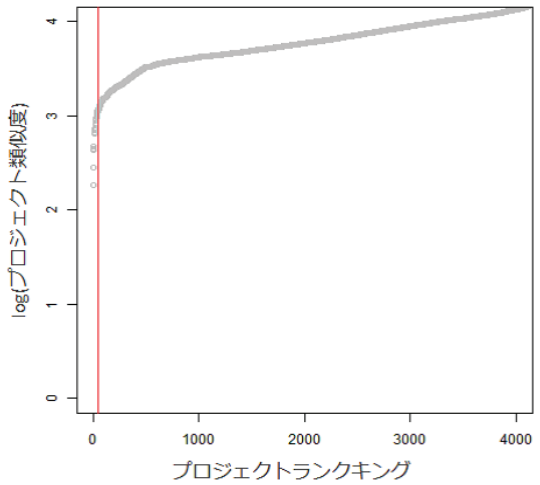

図 3 プロジェクト類似度算出のイメージ ($\theta = 45$)

机上検証での評価は次の手順で行った．まず，検索対象の N 件のプロジェクトから任意のプロジェクトを 1 件抽出する．抽出したプロジェクトのプロジェクト属性を利用して検索を実行する．この際「重要視フラグ」は設定しない．検索結果から得られた類似プロジェクトの生産性の統計量 25 パーセンタイルと 75 パーセンタイルの範囲に，抽出したプロジェクトの生産性が入っていれば「的中した」とみなす．

上記の試行を本検索機能と AND 検索の 2 種類に対して行った．AND 検索では検索条件数を 1〜8 とし，取りうる全ての検索条件パターンを用いた．開発規模は定量的データであるため，検索値から ±50%以内の値であれば，条件と一致しているとみなした．

115 プロジェクトを対象に試行を実施し，検索結果が 0 件であった回数（ミスヒット数）と検索結果が的中した回数（的中回数）をカウントした．さらに，検索結果が 1 件以上であった試行数に対する的中回数を「的中率」とした．ミスヒット数は，ユーザが類似プロジェクトの情報を全く得られなかった頻度を表している．的中率は，ユーザが生産性の妥当性を検証するのに際し，検索結果が妥当なものであったかを表している．ミスヒット数を低減し的中率を向上させることで，ユーザが生産性の妥当性検証をするのに必要な質・量の情報を提示できるようになると考えられる．

表 9 に本検索機能，および AND 検索を利用した結果を示す．表 9 より本検索機能と AND 検索（8 条件）を比較すると，本検索機能を用いることでミスヒット率が減少し，的中率も 39.5 ポイント増加していることがわかる．また，本検索機能と AND 検索（1〜8 条件合計）の的中率を比較すると，本検索機能と AND 検索（1〜8 条件合計）では，大きな差がないことがわかる．

表 10 に AND 検索において検索結果が的中した際に使用されていたプロジェクト属性の頻度を示す．表 10 からは，的中した場合に使用されたプロジェクト属性の頻度に大きな差がないことが確認できる．そのため，AND 検索においてヒット率および的中率を向上させるようなプロジェクト属性には偏りが存在しないことがわかる．

上記の結果より，AND 検索のみしか使えない場合，ユーザは類似プロジェクトを探すために試行錯誤が必要であることが示唆される．これに対し，提案するシステムを用いることで，ユーザは 1 回の検索で類似プロジェクトを探すことが可能となる．これより効率的に類似プロジェクトを検索することが可能となる．

使用感の調査

本検索機能の使用感について β 版の利用者 9 名へアンケート調査を行った．いずれの回答者ともにプロジェクト管理に関する社内資格を保有している．アンケートで

表 9 AND 検索との比較結果

利用した検索機能	試行回数	ミスヒット数	的中回数	的中率
本検索機能	115	0	72	62.6%
AND 検索（8 条件）	115	37	18	23.1%
AND 検索（1〜8 条件合計）	29325	1372	17736	60.5%

表 10 的中時に使用されたプロジェクト属性の頻度

検索条件数	案件区分	顧客区分	サービス種別	基盤区分	開発区分	言語	顧客業種	開発規模
1	100%	100%	100%	100%	100%	100%	100%	100%
2	86.7%	85.6%	87.3%	83.9%	72.4%	84.3%	83.1%	70.3%
3	80.0%	77.5%	78.6%	74.6%	67.5%	73.7%	71.1%	62.4%
4	68.4%	65.6%	67.2%	61.8%	57.7%	61.3%	58.5%	51.0%
5	54.4%	52.6%	53.5%	48.3%	46.7%	48.8%	45.8%	39.5%
6	39.1%	38.3%	38.9%	34.8%	35.5%	36.3%	33.7%	28.9%
7	25.6%	25.3%	25.6%	23.5%	25.1%	25.2%	23.4%	21.1%
8	15.7%	15.7%	15.7%	15.7%	15.7%	15.7%	15.7%	15.7%

は「想定する利用シーンにおいて本検索機能は役立つと思うか」という選択形式の設問に対し，「思う」「ややそう思う」が 9 件，「あまりそう思わない」「思わない」は 0 件であった．

上記を選択した理由（複数回答可能）については，

- 1 度の検索で多くの類似プロジェクトの情報を閲覧することが可能なったため：9 件
- 完全に検索条件に一致しているプロジェクトのみならず，類似プロジェクトを参考にできるため：6 件
- 自分の感覚に合った類似プロジェクトが多く表示されているため：3 件

であった．このことから，本検索機能において想定したユースケースが達成されていると考えられる．

一方で「自身が想定していた結果と異なるプロジェクトが表示されたため少し違和感がある」，「検索に利用できるプロジェクト属性を追加してほしい」，「プロジェクト属性としてステークホルダーの数や調達方法があるとよい」といった改善に関する意見も得た．

5. 関連研究

クラスタリング [13] は個体の集合であるデータセットをクラスタと呼ばれる集合に分割する手法である．分割時は個体同士に定義された類似度に基づき，類似する個体同士を自動的に分割するため，教師データを必要としない．代表的な手法として，最短距離法などの階層的手法，k-means 法 [6] などの非階層的手法に分けられる．本研究においても，類似プロジェクトを探す際に，階層的手法による分割結果を利用するなどといった応用可能性がある．また近年では，k-匿名性への応用に関する研究 [3] が進められている．

今回の検索システムで利用したプロジェクト属性については，日本国内の複数企業を対象にその実績データ収集が進んでいる．例えば，情報処理推進機構 ソフトウェア高信頼化センター（IPA/SEC）によるソフトウェア開発データ白書 [10] や日本情報システム・ユーザー協会（JUAS）によるソフトウェアメトリックス調査 [9] が挙げられる．上記のように，プロジェクト属性については多くの組織で収集活動が行われている．そのため本研究で提案するプロジェクト類似度は他組織で適用できると考えられる．

開発プロジェクトの類似性に着目した研究としては，Azzeh らの研究 [1] や Bjarnason らの研究 [2] のように距離モデルに基づくものが多い．このようにプロジェクトの類似性に着目した研究の用途として，工数見積もりへの適用が多く存在する [4], [8]．このような工数見積もり手法では「類似するプロジェクトにおいては工数も類似する」という仮定の下，プロジェクト属性が似通ったプロジェクトを検索し，工数を予測する．類似度の算出にあたっては本研究と同様に距離に基づくモデルを採用しているものも多い．本研究では類似するプロジェクトに関して参照する情報を，工数以外の品質や開発成果物といった情報にも着目し，見積もり以外の用途も対象としている．

6. おわりに

本稿では NTT データにおける類似プロジェクト検索技術の開発の取り組みについて紹介した．まず類似プロジェクトの情報利用シーンごとに，プロジェクトの類似性を判断する際の観点を調査した．調査結果を参考に，プロジェクト類似度を定義し，プロジェクト類似度に基づく類似プロジェクト検索機能を開発した．本検索機能は 2018 年 4 月に全社およびグループ会社向けに β 版公開されている．

今後の課題として，下記の 2 点が挙げられる．

重み付け手法の改善 今回採用したプロジェクト属性の重みは，アンケート結果をもとにヒューリスティックに決定している．今後，類似プロジェクト検索機能を運用する中で収集した，ユーザが指定した検索条件をもとにオンライン学習を行うことで自動的に重みを更新する仕組みを検討する．

熟練者が持つ知見の形式知化　今回は特定のユースケースを対象に評価を行った．ユースケースにより，検索結果の利用方法や必要なプロジェクト属性は異なると考えられる．熟練者へのヒアリングなどにより，今後も継続的にプロジェクト属性の追加検討を行う必要がある．

謝辞　プロジェクト類似度の検討にあたっては，NTTデータ数理システムの平野直人氏にコメントを頂いた．

参考文献

[1] Azzeh, M., Neagu, D., and Cowling, P. Software project similarity measurement based on fuzzy C-means, *Proc. ICSP'08*, pp. 123-134 (2008).
[2] Bjarnason, E., Smolander, K., Engström, E., and Runeson, P. : A theory of distances in software engineering, *Information and Software Technology*, Vol. 70, pp.204-219 (2016).
[3] Byun, J., Kamra, A., Bertino, E., and Li, N. : Efficient k-anonymization using clustering techniques, *Proc. DASFAA'07*, pp.188-200 (2007).
[4] Keung, J. W., Kitchenham, B. A., and Jeffery, D. R.: Analogy-X: Providing statistical inference to analogy-based software cost estimation, *IEEE Trans. Softw. Eng.*, vol. 34, no. 4, pp. 471-484 (2008).
[5] Kitchenham, B. and Mendes, E.: Why comparative effort prediction studies may be invalid, *Proc. PROMISE2009* (2009).
[6] MacQueen, J. : Some methods for classification and analysis of multivariate observations, *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, pp.281-297 (1967).
[7] Project Management Institute: A Guide to the Project Management Body of Knowledge(PMBOK GUIDE) - Sixth Edition, Project Management Institute (2017).
[8] Shepperd M. and Schofield, C.: Estimating software project effort using analogies, *IEEE Trans. Softw. Eng.*, vol. 23, no. 11, pp. 736-743 (1997).
[9] 一般社団法人日本情報システム・ユーザー協会：ユーザー企業ソフトウェアメトリックス調査 2016 ソフトウェア開発・保守・運用の評価指標, 一般社団法人日本情報システム・ユーザー協会 (2016).
[10] 独立行政法人情報処理推進機構 技術本部ソフトウェア高信頼化センター：ソフトウェア開発データ白書 2016-2017，独立行政法人情報処理推進機構 (2016).
[11] 中村英恵，佐藤慎一，藤江　宏，端山　毅：多様なプロジェクト特性を考慮した品質・生産性実績データ提供システムの構築，ソフトウェア品質シンポジウム 2011 予稿集 (2011).
[12] 伏田享平: NTT データにおける開発データの収集・分析に関する取り組み, ウィンターワークショップ 2016・イン・逗子 論文集, pp.59-60 (2016).
[13] 宮本定明：クラスター分析入門ーファジィクラスタリングの理論と応用，森北出版 (1999).