2007－SE－155 (9)
2007／3／22

相関ルールマイニングの適用によるソフトウェア生産性の決定要因の分析

浜野 康裕†, 水野 修†, 菊野 亨†, 菊地 奈穂美‡, 平山 雅之‡
†大阪大学大学院情報科学研究科
‡独立行政法人 情報処理推進機構 ソフトウェア・エンジニアリング・センター

概 要

ソフトウェア生産性を高めることは，企業の目的である利潤という点において，直接的に影響を与える重要な要素である．従来研究ではプロジェクトデータに回帰分析を適用して生産性を決定する要因を明らかにしている．しかし，この要因は開発環境に強く依存するため，他の企業群に適用することは難しい．しかも回帰分析では，一部にでも欠損するデータ項目が含まれると適用が出来なくなる．本論文では，国内企業15社で実際に行われたプロジェクトから収集されたデータに対して，相関ルールマイニング手法を適用し，生産性と生産性決定要因の間の関係を明らかにした．その結果，生産性が良くなる3つのパターンと生産性が悪くなる2つのパターンを明らかにした．

Software Productivity Analysis Using Association Rules Mining

Yasuhiro Hamano†, Osamu Mizuno†, Tohru Kikuno†, Nahomi Kikuchi‡, Masayuki Hirayama‡
†Graduate School of Information Science and Technology, Osaka University.
‡Software Engineering Center, Information-technology Promotion Agency.

Abstract

Clearly high productivity in software development brings a big profit to companies.In this paper we try to investigate important factors to raise produtivity by analyzing project data which include missing values. By now several studies tried to clarify the important factors by applying regression analysis to project data. But the obtained factors are strongly dependent on development processes. Addionally regression analysis cannot apply to project data with missing values. Thus we newly apply association rules mining technique to project data. By analyzing resultant association rules with respect to productivity, we get 14 important factors. Then we find several patterns in association rules which are essential to high productivity.

1 まえがき

企業の競争力を確保し，開発プロジェクトから高い利潤を得るためには，ソフトウェア生産性を改善することが重要である．生産性の決定要因に関する分析は様々な研究がある [2, 4]．しかし，生産性に影響を与える要因は個々の開発環境に依存するため，ある開発現場での要因がそのまま他の開発現場にも必ずしも適用できるわけではない．つまり，ある開発現場の生産性要因を知るためには，実際にその開発現場から取得されたデータを分析することが賢明である．

ソフトウェア開発プロジェクトデータから生産性決定要因を抽出する手法としては，回帰分析がある [4]．しかし，回帰分析手法では欠損値が1つでも含まれているデータは扱えないため，欠損値が多いデータには適用できない問題があった．そこで，本研究では生産性決定要因抽出のために，相関ルールマイニング手法を用いた．相関ルールマイニングとはデータマイニング手法の1種であり，特に事象間の強い関係をルールとして抽出する手法である [3,5–8]．本手法では，データの一部に欠損を含んでいるデータ

からも，生産性に関する有益な情報を抽出することができる．また，単に生産性に強い影響を与える要因を抽出できるだけではなく，要因の組み合わせに関する情報も得ることが可能となる．これによりより生産性の良否を判断するための有益な材料を多く得ることができた．

今回分析対象としたデータは，独立行政法人 情報処理推進機構ソフトウェア・エンジニアリング・センターが定義し，日本企業15社から収集したエンタプライズ系ソフトウェア開発プロジェクトのデータ1009件である [9]．このデータから生産性を導出し，相関ルールマイニングを適用した結果，生産性の良否を決定付ける要因のいくつかのパターンを発見することができた．

本論文の以降の構成は以下の通りである．まず2節では，対象データと相関ルールマイニングについて述べる．次に3節では，対象データの生産性の導出から，データ項目の前処理，相関ルールマイニングを適用しルールを抽出するまでの流れを述べる．そして4節では，考察として得られた結果の分析を行う．最後に5節で，まとめと今後の課題について述べる．

2 準備

2.1 対象とデータ

今回の分析の対象となるデータは，独立行政法人情報処理推進機構ソフトウェア・エンジニアリング・センター (以下IPA/SECと呼ぶ) が定義し，日本企業15社からIPA/SECへ収集したエンタプライズ系ソフトウェア開発プロジェクトのデータ1009件である (以下SECデータと呼ぶ) [9]．収集されたデータは，プロジェクト特性 (開発種別，業種，業務，アーキテクチャ，主開発言語など)，規模，工期，工数などの情報を含んでいる．

また収集されたデータは，提供企業の品質保証部門や生産部門で精査されたものである．さらに，IPA/SEC側で受領した後にも精査を繰り返し，異常値や誤記のチェックが行われてデータの信頼度が確保されている．

2.2 相関ルールマイニング

相関ルールマイニングは，相関ルール (以下ルールと呼ぶ) と呼ばれる事象間の強い関係を知識として発見する分析手法である [1]．

データ集合全体を分析して (つまり，データのマイニングを実施して)，「ある事象 A が発生するならば別の事象 B も発生する」という事実を発見し，それをルールとして抽出する．このとき，抽出されたルールを $A \Longrightarrow B$ と表記し，A を前提，B を結論と呼ぶ．

このルールの重要度を評価するパラメータとして，「信頼度 (confidence)」と「支持度 (support)」の2つがある．まず，信頼度とは事象 A が発生した場合に事象 B も同時に発生する確率 ($p(B \mid A)$) を表す．つまり，この値が1に近づくほど，ルールの前提と結論の結び付きが強いことを意味する．また，支持度とはルールの出現頻度を表すもので，データ集合全体の中で A と B が同時に発生する確率 ($p(A \wedge B)$) である．

実際のルール抽出では最低信頼度と最低支持度を設定し，その条件を満たすルールだけを抽出する．

3 生産性決定ルールの抽出手法

本節では，生産性決定ルールの抽出手法の詳細な説明を行う．抽出手順の大まかな流れを図1に示す．

3.1 フェーズ1：生産性の計算と分析対象プロジェクトの抽出

本研究では，生産性 P を以下の様に定義した．

$$\text{生産性 } P = log(FP \text{ 実測値}/\text{工数})$$

FP 実測値 とはファンクションポイントを計測した値である．工数 は5つの開発工程 (基本設計，詳細

表 1: 分析対象データの生産性区分

生産性区分	データ件数	生産性の値
生産性が良い	8 件	$P \geq 0.529114$
どちらでもない	34 件	$-0.73392 < P < 0.529114$
生産性が悪い	9 件	$P \leq -0.73392$

設計，製作，結合テスト，総合テスト) の実績工数の和と定義した．

SEC データ 1009 件のプロジェクトのそれぞれについて，生産性 P を計算するために，FP 実測値，及び 5 つの開発工程の実績工数の合計 6 個のデータ項目について，すべての回答が存在し，しかも回答の値が 0 では無いプロジェクトのみを分析の対象とした．以上の手続きによりプロジェクトデータを 51 件抽出した．

51 件のデータについて生産性 P の良否を判定する基準として，「生産性 $P \leq 51$ 件の P の平均 − 51 件の P の標準偏差」ならば「生産性が悪い」とした．一方，「生産性 $P \geq 51$ 件の P の平均 + 51 件の P の標準偏差」ならば「生産性が良い」とした．それ以外のものを「どちらでもない」とした．その結果，51 件のデータは表 1 の 3 つに分類された．

3.2 フェーズ 2：データ項目の絞込みと候補データの前処理

3.2.1 要因候補の選択

SEC データには 200 以上のデータ項目が存在する．IPA/SEC との協議の中で，著しく欠損率が高いデータ項目を除き，さらに生産性決定要因候補を検討した結果，データ項目を 41 項目に絞った．絞り込んだ結果のデータ項目にも欠損値は含まれている．以下では，データ項目は Q_x の形で表記する．x はソフトウェア開発データ白書 [9] における項目番号である．選択された 41 項目の一覧を表 2 に示す．

以下の処理では，51 件のプロジェクト，及び 41 種のデータ項目と生産性良否に対して，生産性の決定要因に関するルールの抽出を行う．

3.2.2 候補データの前処理

本論文で用いる手法である相関ルールマイニングでは，名義的なデータのみを対象としている．そのため順序データを意味的に扱うことはできない．またデータ項目中で著しく数が少ない回答は，最低支持度により手法の過程で破棄されるため，ルールとして抽出することが困難である．そこで分析の前処理として，対象データの再グループ化を行った．実際には以下の二種類の操作を行った．

- 数値データを値の大小により 4 つのグループに分割
- カテゴリカルデータの再グループ化

数値データとしては，「Q_10128：プロジェクト全体の実績月数」のみが対象に含まれている．このデータ項目の有効回答は 0.5ヶ月から 14.6ヶ月までの 43 件が存在している．これを回答の期間により 4 等分に分割した．

S 4ヶ月未満：11 件

M 4ヶ月以上 6ヶ月未満：10 件

L 6ヶ月以上 8.5ヶ月未満：11 件

LL 8.5ヶ月以上：11 件

カテゴリカルデータの再グループ化を行ったのは表 3 の 12 個のデータ項目である．本論文では代表例として「Q_501:要求仕様の明確さ」の変換について説明する (図 2)．「Q_501:要求仕様の明確さ」の回答は，「非常に明確」，「明確」，「やや不明確」，「不明確」の 4 段階存在している．これを「非常に明確あるいは明確」と「やや不明確あるいは明確」に再グループ化した．これにより 4 つの回答が 2 つになった．

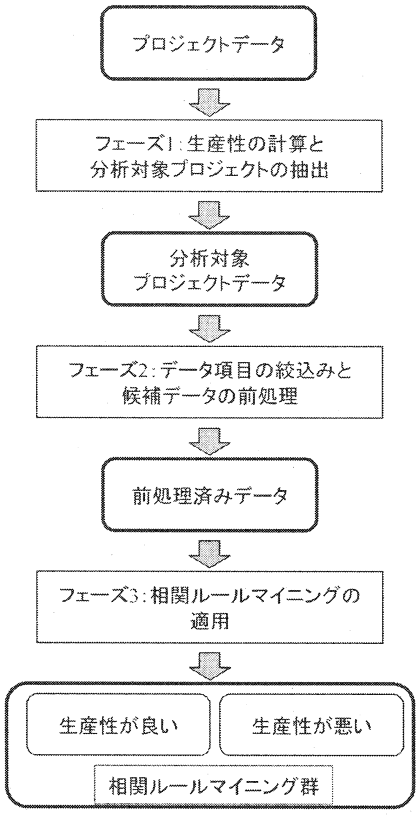

図 1: ルール抽出手法の流れ

表 2: 生産性決定要因候補

分類	項目番号	データ項目
開発プロジェクト全般	Q_103	開発プロジェクト種別
	Q_104	母体システムの安定度
	Q_105	開発プロジェクトの形態
	Q_108	新規の顧客か否か
	Q_109	新規の業種・業務か否か
	Q_110	新規協力会社か否か
	Q_111	新技術を利用する開発か否か
	Q_112	開発プロジェクトチーム内での役割分担・責任所在の明確さ
	Q_113	達成目標と優先度の明確さ
	Q_114	作業スペース
	Q_116	プロジェクト成否に対する自己評価
	Q_117	顧客満足度に対する主観評価
利用局面	Q_204	利用形態
システム特性	Q_301	システムの種別
	Q_302	業務パッケージ利用の有無
	Q_307	処理形態
	Q_308	アーキテクチャ
	Q_309	開発対象プラットフォーム
	Q_312	主開発言語
開発の進め方	Q_401	開発ライフサイクルモデル
	Q_403	類似プロジェクトの参照の有無
	Q_404	プロジェクト管理ツールの利用
	Q_405	構成管理ツールの利用
	Q_406	設計支援ツールの利用
	Q_407	ドキュメント作成ツールの利用
	Q_408	デバッグ・テストツールの利用
ユーザ要求管理	Q_501	要求仕様の明確さ
	Q_502	ユーザ担当者の要求仕様関与
	Q_503	ユーザ担当者のシステム経験
	Q_509	ユーザ担当者の受け入れ試験関与
	Q_512	要求レベル（信頼性）
	Q_514	要求レベル（性能・効率性）
	Q_518	要求レベル（セキュリティ）
	Q_519	法的規制の有無
要因等スキル	Q_601	PM スキル
	Q_602	業務分野の経験
	Q_603	分析・設計経験
	Q_604	言語・ツール利用経験
	Q_605	開発プラットフォームの使用経験
品質信頼性	Q_1010	テスト体制
	Q_1011	定量的な出荷品質基準の有無
工期	Q_10128	プロジェクト全体の実績月数

3.3 フェーズ 3：相関ルールマイニングの適用

3.2 のデータに対して，相関ルールマイニング手法を適用する．本研究では生産性の決定要因を探るため，結論が「生産性が良い」あるいは「生産性が悪い」となっているルールのみを抽出する．すなわち，以下の形で表されるようなルールのみを抽出する．

- $Q_{x1} \wedge \cdots \wedge Q_{xn} \Longrightarrow$ 生産性が良い
- $Q_{y1} \wedge \cdots \wedge Q_{ym} \Longrightarrow$ 生産性が悪い

3.4 適用結果

まず最低支持度を 0.05 に，最低信頼度を 0.80 としてルール抽出を行なった．その後引き続いて，生産性が悪いルールと良いルールのそれぞれに対して，支持度，信頼度が共に高いルールを上位から数個を抜き出した．その結果，生産性が良いルールが 19 個，悪いルールが 11 個求まった．その詳細を表 4 と表 5 にそれぞれ示している．表 4 について説明する．この表の 1 行目は次の 4 つのデータ項目から構成されている．

- 母体システム：非常に安定 (Q_104=a)
- チーム内での役割分担・責任所在：非

表 3: 再グループ化対象項目

項目番号	データ項目
Q_113	達成目標と優先度の明確さ
Q_114	作業スペース
Q_301	システムの種別
Q_309	開発対象プラットフォーム
Q_312	主開発言語
Q_501	要求仕様の明確さ
Q_502	ユーザ担当者の要求仕様関与
Q_503	ユーザ担当者のシステム経験
Q_509	ユーザ担当者の受け入れ試験関与
Q_512	要求レベル (信頼性)
Q_514	要求レベル (性能・効率性)
Q_518	要求レベル (セキュリティ)

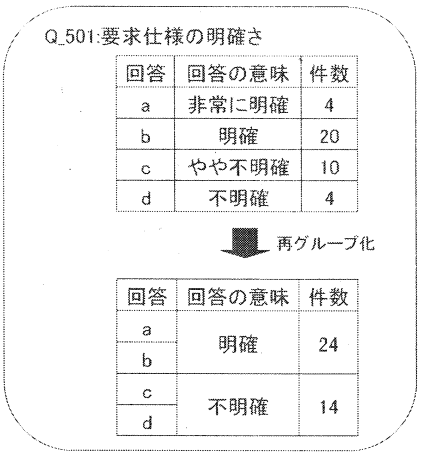

図 2: データ項目の再グループ化の一例

常に明確 (Q_112=a)

- システム種別:アプリケーションソフト (Q_301=a)
- 信頼性の要求レベル:低い (Q_512=cd)

以上の 4 つの条件がデータ中の 0.078 の割合で同時に発生しており，その条件が成り立つ場合は 1 の確率で生産性が良くなることを示すルールである.

4 考察

4.1 得られたルールに含まれる要因

生産性の良否それぞれについて得られたルールに含まれていた要因を列挙する．生産性が良いルールに含まれていた要因は以下の 14 個である.

- 母体システム:非常に安定 (Q_104=a)
- 開発形態:受託開発 (Q_105=b)
- チーム内での役割分担・責任所在:非常に明確 (Q_112=a)
- システム種別:アプリケーションソフト (Q_301=a)
- 開発ライフサイクルモデル:ウォーターフォール (Q_401=a)
- プロジェクト管理ツールの利用:無し (Q_404=b)
- 設計ツールの利用:無し (Q_406=b)
- 要求仕様の明確さ:明確 (Q_501=ab)
- ユーザ担当者の要求仕様関与:十分 (Q_502=ab)
- ユーザ担当者のシステム経験:十分 (Q_503=ab)
- ユーザ担当者の受け入れ試験関与:不十分 (Q_509=cd)
- 要求レベル (信頼性):低い (Q_512=cd)
- 要求レベル (セキュリティ):低い (Q_518=cd)
- プロジェクト全体の実績月数:8.5ヶ月以上で最も長い分類 (Q_10128=LL)

一方，生産性が悪い方に含まれていた要因は以下の 10 個である.

- 開発形態:受託開発 (Q_105=b)

- チーム内での役割分担・責任所在：明確 (Q_112=b)
- 個人の作業スペース:狭い (Q_114=cd)
- 処理形態：対話処理 (Q_307=b)
- 開発ライフサイクルモデル:ウォーターフォール (Q_401=a)
- 設計ツールの利用：無し (Q_406=b)
- ドキュメント作成ツールの利用：有り (Q_407=a)
- デバッグツールの利用：無し (Q_408=b)
- ユーザ担当者の要求仕様関与：十分 (Q_502=ab)
- 法的規制の有無：無し (Q_519=c)

4.2 良否間の要因比較

4.1 で列挙した要因の比較を行う.
良否の両方に現れている要因としては,

- 開発形態：受託開発 (Q_105=b)
- 開発ライフサイクルモデル:ウォーターフォール (Q_401=a)
- 設計ツールの利用：無し (Q_406=b)
- ユーザ担当者の要求仕様への関与：十分 (Q_502=ab)

の 4 つが挙げられる. これらの要因は, 生産性を良くも悪くもする可能性があることが分かった.

良否双方の結果に同じデータ項目が異なる回答で現れている要因としては「Q_112：チーム内の役割分担・責任所在」が挙げられる.「Q_112：チーム内の役割分担・責任所在」が「非常に明確」であれば生産性が良くなり,「明確」ならば生産性が悪くなる結果になった.「非常に明確」と「明確」は共に役割分担・責任所在の点では肯定的な回答に見える. しかし, 本研究の分析対象データ 51 件の回答分布は以下の様になっている.

- 非常に明確：13 件
- 明確：26 件
- 不明確：1 件
- 欠損値：11 件

欠損値を除いた 40 件のうち, 39/40 の 98% が「非常に明確」か「明確」と回答されている. すなわち, 生産性にとっては「役割分担:明確」は必ずしも肯定的な事象ではないと判断できる. よって, 生産性の良否双方に唯一背反の事象が含まれている「Q_112：チーム内の役割分担・責任所在」は, 生産性良否にとってもっとも特徴的な要因として抽出されたと考える.

4.3 特徴的な組み合わせ

生産性が良いルールが 19 個, 悪いルールが 11 個求まったと 3.4 で述べた. 個々のルールを見てみると, 複数ルールに出現する要因項目がある一方で, ごく少数ルールにのみ出現する要因項目がある. また, 組み合わせで出現している要因項目もある. そのため, ルールに出現している要因項目のパターンを見出すことを考えた. パターンが抽出できると, 主要なパターンのルールには単独で意味のある要因項目だけでなく, 組み合わせで意味をなす要因項目の組を示されることから, 本研究で目的としている生産性の決定要因として有用性があると考えたためである.

ルールの要因項目の一覧を作成してみたところ, 表 4 と表 5 のようになった. 得られたルールが含まれる要因の傾向により, いくつかのグループに分類すると次のようになる. 生産性が良いルール (表 4) は, Good_Rule 1 から 9, Good_Rule 10 から 12, Good_Rule 13 から 19 の 3 つに分類した. 生産性が悪いルール (表 5) は, Bad_Rule 1 から 3 と Bad_Rule 4 から 11 の二つに分類した. 最終的な分類と含まれる要因の意味を表 6 に示す. 生産性が良い方では 3 つのパターン, 生産性が悪い方では 2 つのパターンが存在していることが分かった.

表 4: 結果 1(生産性が良い場合:支持度 0.078，信頼度=1)

	Q_104 =a	Q_105 =b	Q_112 =a	Q_301 =a	Q_401 =a	Q_404 =b	Q_406 =b	Q_501 =ab	Q_502 =ab	Q_503 =ab	Q_509 =cd	Q_512 =cd	Q_518 =cd	Q_10128 =LL
Good_Rule1	o		o	o								o		
Good_Rule2	o		o		o							o		
Good_Rule3	o		o			o						o		
Good_Rule4	o		o				o					o		
Good_Rule5	o		o					o				o		
Good_Rule6	o		o						o			o		
Good_Rule7	o		o							o		o		
Good_Rule8	o		o									o	o	
Good_Rule9	o	o	o									o		
Good_Rule10	o		o								o	o		
Good_Rule11			o					o			o	o		
Good_Rule12			o						o		o	o		
Good_Rule13		o											o	o
Good_Rule14		o		o									o	o
Good_Rule15		o			o								o	o
Good_Rule16		o					o						o	o
Good_Rule17		o							o				o	o
Good_Rule18		o								o			o	o
Good_Rule19		o										o	o	o

表 5: 結果 2(生産性が悪い場合:支持度=0.137，信頼度=1)

	Q_105 =b	Q_112 =b	Q_114 =cd	Q_307 =b	Q_401 =a	Q_406 =b	Q_407 =a	Q_408 =b	Q_502 =ab	Q_519 =c
Bad_Rule1		o	o	o			o			
Bad_Rule2		o	o	o				o		
Bad_Rule3		o	o	o					o	
Bad_Rule4	o	o	o							
Bad_Rule5	o	o	o	o						
Bad_Rule6	o	o	o		o					
Bad_Rule7	o	o	o			o				
Bad_Rule8	o	o	o				o			
Bad_Rule9	o	o	o					o		
Bad_Rule10	o	o	o						o	
Bad_Rule11	o	o	o							o

5 まとめ

本研究では，国内企業 15 社で実際に行われたプロジェクトから収集されたデータに対して，相関ルールマイニング手法を適用し，生産性と生産性決定要因の間の関係を明らかにした．その結果，生産性決定要因として，生産性を良くする 3 つのパターンと，生産性を悪くする 2 つのパターンを発見することができた．特に「Q_112：チーム内の役割分担・責任所在」は生産性にとってもっとも特徴的な要因として抽出された．

今後の課題としては，本研究で得られたパターンを別のデータに適用することで，生産性の予測を行い，得られた結果の妥当性を検証することが考えられる．

謝辞

本研究は平成 18 年度三菱総合研究所委託研究「プロジェクトのメトリクスを活用したリスク抽出・管理・予測手法の検討・実適用・評価」から援助を受けて実施された．

表 6: 生産性決定要因

パターン	項目名	選択肢	データ番号と選択肢記号
良いパターン 1	母体システム	非常に安定	Q_104=a
	チーム内での役割分担・責任所在	非常に明確	Q_112=a
	要求レベル (信頼性)	低い	Q_512=cd
良いパターン 2	チーム内での役割分担・責任所在	非常に明確	Q_112=a
	ユーザ担当者の受け入れ試験関与	不十分	Q_509=cd
	要求レベル (信頼性)	低い	Q_512=cd
良いパターン 3	開発形態	受託開発	Q_105=b
	セキュリティ要求レベル	低い	Q_518=cd
	プロジェクト全体の実績月数	最も長い分類 (8.5ヶ月以上)	Q_10128=LL
悪いパターン 1	開発形態	受託開発	Q_105=b
	チーム内での役割分担・責任所在	明確	Q_112=b
	作業スペース	狭い	Q_114=cd
悪いパターン 2	チーム内での役割分担・責任所在	明確	Q_112=b
	作業スペース	狭い	Q_114=cd
	処理形態	対話処理	Q_307=b

参考文献

[1] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[2] S. Han and D. W.Halpin. The use of simulation for productivity estimation based on multiple regression analysis. In *Proc. of the 37th Conference on Winter simulation*, pp. 1492–1499, 2005.

[3] X. Huang, A. An, and N. Cercone. Comparison of interestingness function for learning web usage patterns. In *Proceedings of the 2002 ACM CIKM International Conference on Information and Knowledge Management*, pp. 617–620, 2002.

[4] J.Vousburgh, B.Curtis, R.Wolverton, B.albert, H.Malec, S.Hoben, and Y.Liu. Productivity factors and programming environment. In *Proc. of the 7th International conference on Software engineering*, pp. 143–152, 1984.

[5] S. K. Madria, C. Raymond, S. Bhowmick, and M. Mohania. Association rules for web data mining in whoweda. In *Proceedings of International Conference on Digital Libraries*, pp. 227–233, 2000.

[6] M. Shyu, C. Haruechaiyasak, S. Chen, and N. Zhao. Collaborative filtering by mining association rules from user access sequences. In *International Workshop on Challenges in Web Information Retrieval and Integration*, pp. 128–135, 2005.

[7] X. Wang, M. R. Smith, and R. M. Rangayyan. Mammographic information analysis through association-rule mining. In *Electrical and Computer Engineering*, pp. 1495–1498, 2004.

[8] X. Zhu and X. Wu. Mining video association for efficient database management. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pp. 1421–1432, 2003.

[9] 独立行政法人 情報処理推進機構 (IPA). ソフトウェア開発データ白書 2005. 日経 BP 社, 2005.