情報処理学会論文誌 Vol.53 No.2 495–509 (Feb. 2012)

要求獲得のためのオントロジを Webマイニングにより拡充する手法の提案と評価

海谷 治彦[1,a)] 清水 悠太郎[1] 安井 浩貴[1] 海尻 賢二[1] 林 晋平[2] 佐伯 元司[2]

受付日 2011年5月27日, 採録日 2011年11月7日

概要：ソフトウェア技術者がソフトウェアシステムの要求獲得を行うためには，システムを適用する問題領域の知識が必須である．ドメインオントロジ等の問題領域の知識の明示的な記述は，要求獲得結果を完全かつ正当にすることに貢献する．ドメインオントロジ等の利用を想定した要求獲得技法はいくつか提案されている．そして，対象分野に関する文書をまとめたり，当該分野の専門家から情報を抽出したりすることで，ドメインオントロジを作成することはできる．しかし，一般に要求獲得を行う技術者は問題分野の専門家ではないため，分野に特化した情報のみから構成されるドメインオントロジだけでは，要求獲得を漏れなく誤りなく行うことは難しい．本稿では，Web マイニングの技術を用いてドメインオントロジに技術者がドメイン知識を理解するのに有益な知識を追加し，ドメインオントロジを拡充する手法とツールを提案する．提案手法では，まず，ドメインオントロジにすでに含まれる概念を検索語として用いて，当該概念に追加すべき概念の候補群を Web から自動的に収集する．そして，既存概念ごとに，既存概念との関連の深さや，Web 上の文書における出現頻度や分布に基づき，候補群のランク付けを自動的に行う．これらのランク付けに基づき技術者がオントロジの拡充を行う．拡充されたオントロジが要求獲得結果の漏れのなさ，誤りの少なさを改善できることを，比較実験を通して確認して結果も示す．

キーワード：要求獲得，ドメイン知識，オントロジ，ウェブマイニング

A Method and A Tool for Enhancing Domain Knowledge for Requirements Elicitation Using Web Mining

Haruhiko Kaiya[1,a)] Yuutarou Shimizu[1] Hirotaka Yasui[1] Kenji Kaijiri[1]
Shinpei Hayashi[2] Motoshi Saeki[2]

Received: May 27, 2011, Accepted: November 7, 2011

Abstract: Software engineers require knowledge about a problem domain when they elicit requirements for a system about the domain. Explicit descriptions about such knowledge such as domain ontology contribute to eliciting such requirements correctly and completely. Methods for eliciting requirements using ontology have been thus proposed, and such ontology is normally developed based on documents and/or experts in the problem domain. However, it is not easy for engineers to elicit requirements correctly and completely only with such domain ontology because they are not normally experts in the problem domain. In this paper, we propose a method and a tool to enhance domain ontology using Web mining. Our method and the tool help engineers to add additional knowledge suitable for them to understand domain ontology. According to our method, candidates of such additional knowledge are gathered from Web pages using keywords in existing domain ontology. The candidates are then prioritized based on the degree of the relationship between each candidate and existing ontology and on the frequency and the distribution of the candidate over Web pages. Engineers finally add new knowledge to existing ontology out of these prioritized candidates. We also show an experiment and its results for confirming enhanced ontology enables engineers to elicit requirements more completely and correctly than existing ontology does.

Keywords: requirements elicitation, domain knowledge, ontology, Web mining

1. はじめに

ソフトウェアシステムは単独で利用されることはほとんどなく，現実世界の問題を解決するために，ハードウェア，人間，既存の他のシステム等とともに利用される．たとえば，EasyChair 等の国際会議開催支援システムは，論文投稿，査読過程，結果通知等の人間が行う作業，Web サーバや電子メール等の外部システムとの連携が重要である．本稿では類似した問題に関連する事物の集まりをドメインと呼ぶ．要求分析者はあるドメインに関係する問題を解決するシステム構築において，当該ドメインを理解し，考慮する必要がある．ドメインを理解し，考慮するための最も効果的な方法は，ドメイン専門家とともに作業を行うことである．しかし，そのような専門家は一般に多忙なため，頻繁に共同作業が可能とはいえない．加えて，要求分析者がドメインに関する一定の知識を有していなければ，専門家は真面目に相手をしてくれないであろう．よって，要求分析者はドメインに関する知識を自力で可能な限り，獲得しておく必要がある．

ドメイン知識に関する明示的な記述や文書を参照することは，要求分析者がドメイン知識を得る手段の 1 つとして妥当なものであり，ドメインオントロジはそのような記述や文書の典型例である．ほとんどのドメインオントロジは，知識の単位を概念と呼び，概念を節として，概念間の関係を枝とするグラフ構造で記述される．そのようなグラフ構造を参照することで，分析者は当該ドメインの理解を深めることができる．また，推論機構を提供することで，直接に記述してある以上の情報を容易に得られるオントロジもある．このような理由から，ドメインオントロジを利用した要求獲得技法は数多く提案されている [1], [2], [3], [4]．これらの技法の多くは利用するオントロジの品質に大きく依存するにもかかわらず，オントロジをどのように獲得や構築するかについての議論はあまり行われていない．

要求分析のためのドメインオントロジの構築法として典型的なものとして，ドメインの専門家に作成を依頼したり，ドメインに関する文書をまとめたりする方法がある．どちらの場合でも，オントロジは当該ドメインに特化した内容のみで構成されがちであり，ドメインの専門家ではない要求分析者が理解するには難しいことがある．よって，ドメインの専門家ではない要求分析者が，ドメインに特化した知識を理解する助けになる情報を追加することが有効である．たとえば，国際会議開催支援システムの要求獲得においては，Web 上の共同作業を支えるメカニズム（たとえば電子投票や電子掲示板等）の一部が，論文査読過程の一部を理解し，要求を構築する助けになる．しかし，そのような査読過程に特化していないメカニズムは，国際会議開催支援に関するドメインオントロジから漏れている場合もある．また，そのようなメカニズムは特定ドメインに関係なく進歩している場合が多い．たとえば，Twitter 等の技術も国際会議開催支援で利用可能と思われるが，当該ドメインとは関係なく進歩してきたものである．よって，我々は要求分析者の便宜のために，ドメインから独立した資源を用いて，ドメインオントロジの品質，特に要求分析者がオントロジを利用する際の可用性を改善する必要がある．

そこで，本稿ではドメインに特化した概念を補足する概念を追加することで，ドメインオントロジを拡充するための手法を提案する．国際会議開催支援システムの例でいえば，当該システムや業務に関する知識に加えて，それらを支えうるメカニズムや技術に関する知識を，要求分析者が参照できるような支援することである．要求分析者はコンピュータ技術者の一種であるため，このような技術知識は，ドメイン知識を理解する助けになると思われる．追加する補足概念は Web ページから獲得することとする．Web ページ上の情報は比較的すばやく更新されているものもあるため，多数の新しい概念を追加できる可能性がある．詳細は 4 章で説明するが，本手法では，語の共起関係や頻度等の軽量化自然言語処理に基づく情報を用いるため，追加する補足概念候補の収集を自動的に遂行することができる．我々は提案手法を遂行するための支援ツール OREW の構築も行い，OREW を用いた比較実験を通して手法の評価も行った．

本稿の構成は以下のとおりである．次章では，なぜオントロジの拡充が重要であるか，および，どのような概念を拡充することが有効であるかを説明するために，関連研究を概観する．3 章では，ドメインオントロジを利用する要求獲得法の 1 つである ORE 法の紹介をする．次に 4 章においてオントロジ拡充法を説明し，5 章において拡充法の実施を支援するツール OREW を紹介する．6 章において提案するオントロジ拡充法の評価のための実験とその結果を示す．実験から，拡充されたオントロジを用いる場合，拡充前のオントロジを用いた場合よりも，漏れも誤りも少ない要求獲得ができることが分かった．最後にまとめと今後の展望を述べる．

2. 関連研究

本稿ではドメインという言葉を類似した問題に関係する事物の集合としている．ソフトウェアプロダクトラインの分野では，ドメインを識別し再利用することが有益であるとされている [5]．一方，オントロジという語句の最も有名な定義の 1 つは “formal explicit specification of shared conceptualization” [6] である．よって，ドメインオントロ

[1] 信州大学
Shinshu University, Nagano 380–8553, Japan
[2] 東京工業大学
Tokyo Institute of Technology, Meguro, Tokyo 152–8552, Japan
[a)] kaiya@shinshu-u.ac.jp

ジとは，類似する問題に関連する事物の明示的な仕様であると考えてよい．

ソフトウェア工学の分野でオントロジは広く利用されている [7]．たとえば，ドメインオントロジがプログラム理解に利用されている [8]．要求工学分野では，ドメインオントロジを利用した要求獲得技法が数多く研究されている [1], [2], [3], [4]．ドメインオントロジは要求項目の抜けや誤りを分析者が検出するのに利用される場合がある．これらの技法では高品質なドメインオントロジが存在することが前提となっている．しかし，実際には高品質のオントロジを取得したり作成したりすることは容易ではない．それゆえ，高品質なオントロジをどのように作成するかについての研究が必要となる．

要求獲得のためのドメインオントロジを作成するツールとして TCORE [9] が提案されている．TCORE によって，類似した問題領域に関する複数の技術文書（マニュアルや仕様）から，オントロジを作成することができる．しかし，TCORE で作成されたオントロジを用いて要求獲得を行った場合でも，オントロジによって獲得できた要求項目はおよそ 6 割であり，残りは Web ページ等の一般的なリソースに基づき獲得が行われたことが報告された [10]．よって，我々はそのような一般的な情報リソースの有効活用を検討しなければならない．

文献 [11] では，要求獲得で必要な知識を，(a) 問題領域の知識，(b) コンピュータシステムの実現技術や環境に関する知識，(c) 要求仕様書の文書化に関する知識，(d) インタビュやワークショップ等の要求獲得方法に関する知識の 4 通りに分類している．文献 [11] では，(a) と (b) がドメインに特化した知識とされているが，本稿では，(a) のみをドメインに特化した知識とする．なぜならば，問題領域の業務はコンピュータとは無関係に遂行される場合があり，コンピュータに関する知識は，異なる多数のドメインで，共通して利用される場合があるからである．また，本稿で提案する手法の利用は，要求分析者としての知識は保有していることを想定しているため，知識 (c)，(d) に関する知識の拡充は行わない．要求分析者が上記 (a) のような問題領域の知識，たとえば，株式取引や医療事務等を理解するためには，(a) と (b) の知識を関連付けて提供することは有用だと考えられる．なぜならば，要求分析者は技術者の一種であるため，実現技術や環境に関する知識はある程度以上保有しているため，適用可能な技術から，問題領域の概念を理解しやすいからである．たとえば，医療事務における患者の情報は，担当医，技師，事務職員等が異なる権限に基づき，異なるアクセスを行う．このような概念を，要求分析者は，コンピュータシステムのアクセスコントロールの技術と対応付けて，的確に理解することが可能である．

一方，知識を表現する語句の文法的な特徴から，名詞である語句が，ドメインを強く特徴付けているという報告もある [12]．この報告では，名詞のみに着目したほうが，ソフトウェア成果物間のトレーサビリティを高品質に確立することができることが述べられている．この報告における研究自体，ソフトウェアに関する文書等の特定分野の文書では，名詞が他の品詞に比べて，文書の意味の反映していることに基づいている [13], [14]．たとえば，「保存する」，「送付する」，「確認する」，「便利に」，「安全に」，「素早く」等，名詞以外の動詞，副詞等は特定のドメインを特徴付けているとはいえない．一方，「患者」，「カルテ」，「処方箋」等，名詞に相当するものは，特定のドメインを特徴付けているといえる．

一般的なオントロジの作成を支援する研究はいくつか存在し [15], [16], [17]，ほとんどは自然言語処理技術を利用している．ある技法では既存の辞書を利用し [18]，別の技法群では Web 検索（Web Crowling）を用いる [19], [20], [21]．Protege, OntoEdit, KAON, WebODE, TEXT-TO-ONTO 等，オントロジ作成支援ツールは数多く存在し，それらを比較検討した文献も見られる [22], [23]．これらのツールで作成されるドメインオントロジも，上記に示した (a) 問題領域の知識のみを含むものであり，(a) の知識を (b) コンピュータの知識で補う必要がある．

オントロジを拡充する手段の 1 つとして，既存のオントロジを併合する手法群 [24], [25], [26] がある．これらの手法は，同一ドメインのオントロジを併合し，より漏れが少なく，矛盾のないオントロジの構築を目指している．そのため，これらの技法の評価は Precision（精度）と Recall（再現率）を用いて行われている．本稿では，このような包括的なオントロジの構築を目指しているのではなく，要求分析者がドメインオントロジを理解し，利用する補足となる情報を追加することが目的であるため，これら手法群とは目的が異なる．

3. ORE 法について

本章では，ドメインオントロジを利用した要求獲得手法の 1 つである Ontology based Requirements Elicitation（ORE）[3] を紹介する．ORE 法を紹介する理由は以下のとおりである．第 1 に本稿で説明するオントロジ拡充手法の理解を深めるためには，ドメインオントロジを用いた具体的な要求獲得手法を利用するのが適切だからである．なお，本稿で提案するオントロジ拡充手法は ORE 法に特化したものではない．第 2 に，5 章で紹介するオントロジ拡充支援ツールは ORE 法のデータ構造に特化して構築されているため，ツールの説明のために ORE 法に関する解説が必要となる．第 3 に ORE 法に基づくオントロジを用いた要求獲得支援システムが存在するため，そのシステムを用いて拡充されたオントロジの評価を効率的に行うことができるためである．我々は拡充前のオントロジと拡充後のオントロジの利用結果の比較を通して，拡充されたオント

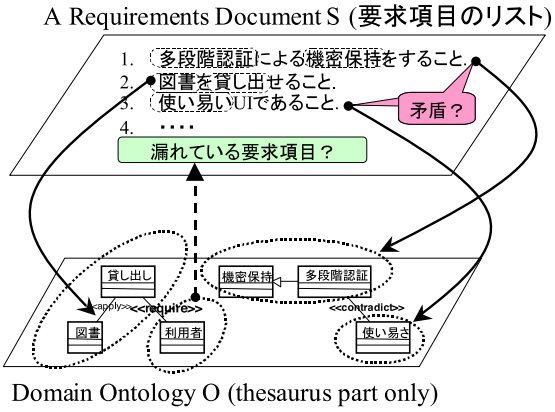


図 1 ORE 法におけるドメインオントロジを利用した要求獲得法
Fig. 1 How to elicit requirements using a domain ontology in ORE.

ロジの評価を行う．もし，支援システムを利用しなければ，拡充されたオントロジによる効果には関係ない因子，たとえば，被験者の違いによる既存の要求項目とオントロジ中の概念との対応付け技能の違い（**図 1** の上下の対応付け）による影響を実験結果から排除しなければならない．支援システムの利用によって，オントロジの拡充の効果には関係ない，被験者の差異から生まれる因子を事前に排除することができるため，評価を効率的に行うことができる．

図 1 に ORE 法を用いた要求獲得において，ドメインオントロジがどのような役割を担うかを説明する．この図は個々の要求仕様書に相当する “A requirements document S” とラベル付けされている上部の層と，ドメインオントロジに相当する “Domain Ontology O” とラベル付けされている下部の層に分かれている．上部の層に列挙された文が個々の要求項目群に相当し，下部の層に描かれたクラス図風のグラフ構造がオントロジに相当する．グラフの節が当該ドメインにおける概念に相当し，その概念を表す語句でラベル付けされている．さらに，それぞれの概念は，その内容によって処理に相当する概念（function），データに相当する概念（object）等の型情報が付記されている．概念間の関係は is-a，has-a，一方が他方を必要とする関係（require），相互に相容れない関係（contradict）等の一般的な関係から，object に function を適用する関係（apply）等，情報システムに特化した関係も定義されている．

要求分析者は初期要求に相当する要求項目群を文書としてステークホルダから獲得しておく．次に，それぞれの要求項目群に含まれる語句を表す概念を探すことで，要求項目と概念を関連付ける．たとえば，図 1 では，要求項目 1 に概念「機密保持」と概念「多段階の認証」に相当する語句が含まれるため，要求項目 1 とこれらの概念が関連付けられている．要求項目と関連付けられた概念と関係がある概念を利用して，不足している要求項目や矛盾した要求項目群を予測することができる．たとえば，概念「図書」と「貸し出し」は既存の要求項目 2 と関連付けられており，概念「利用者」は「貸し出し」から必要（require）とされている．しかし，概念「利用者」に関連する要求項目は存在しない．そこで，ORE 法では，概念「利用者」を含むような要求項目の追加を要求分析者に推薦する．推薦結果と効果については文献 [10] を参照されたい．概念「多段階の認証」は要求項目 1 に関連付けられ，概念「使いやすさ」は要求項目 3 に関連付けられている．オントロジ上にこれら 2 つの概念は相互に矛盾する（contradict）という情報があるため，関連する要求項目の双方もしくは片方を削除もしくは修正する必要性があるかもしれないことを，ORE 法は分析者に助言する．助言結果と効果については文献 [27] を参照されたい．ドメインオントロジに十分な情報が含まれていれば，上記のような要求項目の評価と，評価に基づく分析者による更新を繰り返すことで，より完全かつ正当な要求項目群にすることが可能となる．

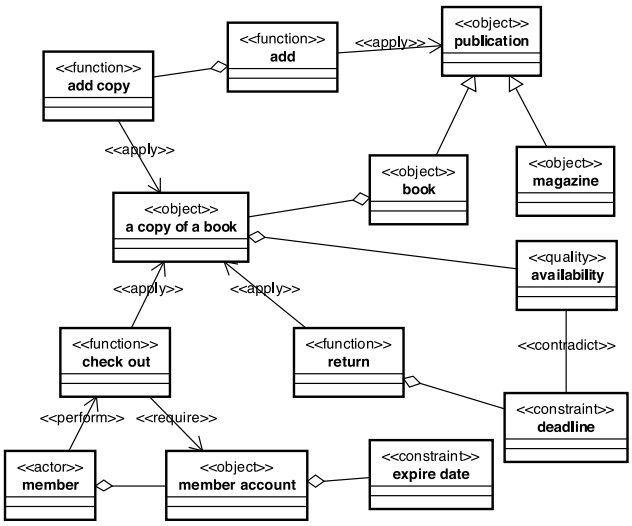


図 2 ORE 法のドメインオントロジの実例（一部）
Fig. 2 A part of an example of an ORE domain ontology.

図 2 に具体的な ORE オントロジの例を示す．この例は図書館等の図書管理システムのためのドメインオントロジである．たとえば，概念 “publication” の一種として “a book” と “magazine” があり，“publication” に対しては “add” という機能（function）を適用（apply）することができる．概念および関連に付記されている function，object，apply，perform 等は型を示しており，情報システムの要求定義に有用な型を事前に準備している [3]．また，型の情報を用いて推論規則を記述することができ，これらの推論規則によって，ORE 法のオントロジは記述されている直接的な関連以上の情報を引き出すことができる．推論規則は以下の例に示すように，Condition と Action の 2 つの部分に分かれており，Condition が成り立つ場合，Action に示すような要求項目の追加，更新を示唆する．この示唆に基づき，ORE 法では要求項目群の更新を要求分析者が行う．

- Conditions:
  - ある要求項目が，オントロジ中の object 型のある概念 O と，function 型のある概念 F1 に対応付く．
    and
  - オントロジ上で，O は F1 以外の function 型の概念 F2 と apply 型の関係で持っている．
- Action:
  - O と F2 に関連付くような新たな要求項目を追加することを示唆する．

既存の要求仕様書に “A copy of a book shall be checked out” という要求項目がすでに存在し，図 2 のオントロジが ORE 法において利用可能だとする．この場合，上記推論規則の O と F1 は，それぞれ “a copy of a book”，“check out” に対応し，F2 は “return” もしくは “add copy” に対応する．そこで，上記推論規則の Action に基づき，“a copy of a book” と “return” が関連付くような新たな要求項目，たとえば “A copy of a book shall be returned” の追加を ORE 法では示唆することができる．

4. オントロジ拡充法

4.1 オントロジ拡充への要求

1 章で述べたように，要求分析者が十分に理解しているドメインに依存しない知識（たとえばコンピュータシステムの知識）とともに，ドメインに特化した知識を提供することで，要求分析者はドメインに特化した知識を理解しやすくなる．たとえば，国際会議開催支援オンラインシステムの場合，論文査読の概念と，オンライン投票システムの概念を関連付けることで，論文査読の概念を要求分析者が理解しやすくなる．しかし，ドメインに依存した知識の理解に貢献しない知識を追加しても意味がない．国際会議開催支援オンラインシステムの例では，オンラインシステムの機能の多くが，TCP（transmission control protocol）を利用して構築されているとしても，TCP 自体の知識はドメイン知識の理解にあまり貢献しない．ドメインに依存しない知識は，ドメインに特化した文書や人物等の資源だけから獲得することは容易でないと思われる．そこで，ドメインに依存しない知識を，ドメインに特化しない資源から収集する手法が必要である．

4.2 拡充法の概要

本拡充法は，既存のオントロジに，要求分析者にとって補足となる情報を追加することが目的である．また，拡充されたオントロジは複数の要求分析者に再利用することを想定している．よって，方法の実施は，分野の専門家と要求分析者が協力して実施するのが望ましい．しかし，要求分析者側が利用することを考えれば，ある要求分析者が分野の専門家の助言を得ながら拡充するのが現実的である．Web 上には多種多様な文書が存在するため，既存のドメインオントロジに含まれる概念を表す語句を用いて検索を行うことで，当該ドメインの理解の助けになる概念を表す語句を収集することが可能である．そこで，我々は 4.1 節の要求にある，ドメインオントロジに追加すべき概念を，Web マイニングによって収集することとした．Web マイニングによるオントロジ拡充法の概要を**図 3** に示す．手法の主たる入力はあるドメインに特化した概念を含むドメインオントロジである．入力は空のオントロジではなく，いくつかの既存概念を含んでいなければならない．このドメインオントロジは図に示す 4 つのステップを踏んで拡充される．それぞれのステップは本章の続きで詳説するが，簡単な概要をここで紹介する．既存概念すべてを検索条件として Web 検索をしても，条件が厳しすぎるため，検索結果が出ない場合がある．そのために，既存概念を意味のあるグループ（サブオントロジ）に分割する（Step 1）．次にそれぞれのサブオントロジごとに検索を行い，Web ページを収集する（Step 2）．収集したページ内のそれぞれの文に着目し，文単位で既存概念と共起関係にある語句を追加すべき概念の候補とする（Step 3）．共起関係にない概念は，既存概念の補足とならないと見なし候補としない．最後に，候補となった概念を既存概念ごとに提示し，対話的に概念を追加する（Step 4）．1 度拡充されたオントロジを再度入力として，拡充を繰り返してもよい．繰り返すか否かは本拡充手法を実施する者が主観的に判断してよい．

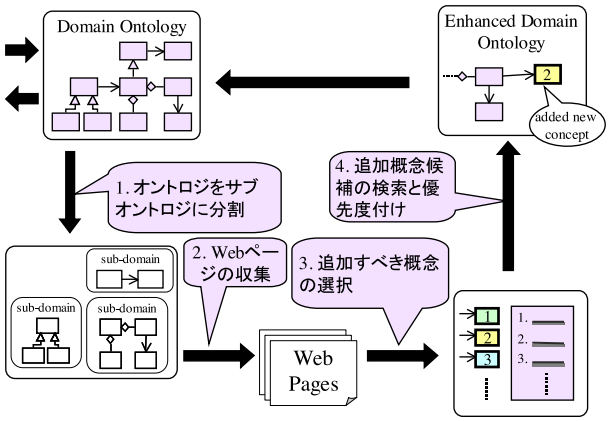


図 3 オントロジ拡充法の概要
Fig. 3 An overview of our ontology enhancement method.

本拡充法では，文中の語句とその品詞，および語句の係り受け関係のみを利用し，複雑な自然言語処理を必要としない．また，品詞および係り受けが存在する自然言語，たとえば英語や日本語であれば適用可能である．

4.3 ステップ 1：オントロジをサブオントロジに分割

あるドメインに特化した入力であるオントロジに含まれる概念の中にも，それほどドメインに特化していない概念もある．たとえば，ある種の機能や品質，制約等を表す概念は多数のドメインに利用される一般的なものである場合

が多い．もし，そのような一般的な概念がWebマイニングに利用されてしまうと，当該のドメインにあまり関係のない概念も検索されてしまう．そこで，我々はこのような一般的な概念をWebマイニングの前に除外する．入力オントロジが3章で述べたORE法に準拠するオントロジの場合，function，quality，constraintの型を持つ概念を自動的に除去する．これらの概念が他のobjectやactor等の概念と比べて一般的であることは，ORE法で利用するオントロジの実例のレビューを通して判断した．レビューの根拠は2章で紹介した，名詞が他の品詞に比べて文書の意味の反映していることに基づいている[13], [14]という文献であり，ORE法におけるobject，actorは名詞に相当し，それ以外は名詞以外に相当するためである．constraintも名詞に相当するが，「二日以上」，「十人以内」等，語句としてドメインを特徴付けることが少ないため除外した．function，qualityの型の概念は，「保存する」，「送付する」，「確認する」，「便利に」，「安全に」，「素早く」等，ドメインを特徴付けているとはいいがたい概念が多かった．

入力オントロジから上記のような一般的なオントロジを除外した後でも，オントロジは多数の概念を保持している場合が多い．それらすべての概念を利用してWebマイニングを行った場合，ほとんど情報を検索できない恐れがある．また，1つのドメインオントロジの中にも，他の概念に比べて結び付きの強いオントロジ群が存在する場合がある．たとえば，銀行システムのドメインオントロジの場合，預金に関連が深い概念群，ローンに関係の深い概念群等が見出せる．我々は1つのドメインオントロジ内にあるこのような概念群をサブドメインオントロジもしくはサブドメインと呼ぶことにする．サブドメインに分割することで，Webマイニングを行った場合，ほとんど情報を検索できないという問題を回避することができる．

我々はオントロジを概念を節とするグラフと見なし，サブドメインを自動的に識別するために，媒介中心性[28]と呼ばれる指標を用いることにした．媒介中心性の値はグラフ中の各節に定義され，直感的にはグラフ中である節を通過する経路が多いほど，媒介中心性の値は大きい．グラフ$G(V,E)$上のある節$v \in V$の媒介中心性の形式的な定義は以下である．

$$BC(v) = \sum_{s,t \in V \ where \ s \neq t} \frac{\sigma_{s,t}(v)}{\sigma_{s,t}}$$

ただし，$\sigma_{s,t}$はsとtの間の最短経路の数であり，$\sigma_{s,t}(v)$はvを通過するsとtの間の最短経路の数である．

我々は一定の値以上の媒介中心性を持つ概念をサブドメインの中心となる概念とすることにした．中心とならないそれぞれの概念cについて，最も距離の近い中心概念が属するサブドメインにcを追加することで，サブドメインを構成することにした．ある概念から複数の中心概念への距離が等しい場合があるため，いくつかの概念は複数のサブドメインに属する場合がある．

4.4 ステップ2：Webページの収集

それぞれのサブドメインについて，Google等のサーチエンジンを用いて，Webページの検索を行う．サブドメインに含まれる概念の名称がすべて含まれていることを検索式とする．ORE法におけるオントロジの場合，図2に示すように，個々の概念はその意味を表現する固有の名称を1個保有する．Webページの収集において，我々は収集されるページの目標値を設定し，収集されたページ数が目標値を上回るか否かに注目する．目標値自体は，サブドメインに含まれる概念の名称がすべて含まれていることを示す検索式を用いて，実際に検索を試行的に行い設定するのがよい．次章での支援ツールでは，この目標値のデフォルト設定は20である．目標値を上回らないということは，Web上にある文書（ページ）の観点から，検索に用いた概念群が相互に無関係すぎることを示している．Webページ収集における第1の問題点は，サブドメインという形である程度相互に関連の深い概念を集めて検索を行っても，検索されるページ数が目標値に達しない場合があることである．第2の問題点は，たとえ目標値以上の数のページが検索されたとしても，ドメインオントロジに概念を追加するという点から不要なページが含まれている場合があることである．これらの問題点をそれぞれ解決するために，我々はシンプソン係数とNormalized Term Frequency（NTF）をそれぞれ用いることとした．

最初に，ある目標値以上のページを検索するためには，検索式を構成する概念名をいくつか除外する必要がある．シンプソン係数は2つの語句の類似性を示す指標の1つである[29]．他の概念名とのシンプソン係数が最小である概念を順番に除外することで，検索されるページ数の増加を目指す．2つの語句ccとcとのシンプソン係数$S(cc,c)$の定義は以下である．

$$S(cc,c) = \frac{Hit(cc\&c)}{Min(Hit(cc), Hit(c))}$$

ここで，$Hit(q)$は語句qが出現するという条件によって検索されるページ数であり，$q\&p$は，語句pとqが双方出現することを示す条件式である．また，$Min(x,y)$はxとyのうちで小さい値を返す．

次に，不要なページを除外するために，それぞれのページのNTFを計算し，NTFがある値より小さいページを不要なページと見なし除外する．NTFは当該のページがサブドメインとどれだけ関係が深いかを示す指標であり，サブドメインオントロジsdのページdのNTFは以下の式に従い計算される．

$$NTF(sd,d) = \frac{\sum_{i=1}^{N} tf(\{d\}, c_i)}{Count(d)}$$

ここで，$Count(d)$ はページ d に含まれる語句の数，N はサブドメインオントロジ sd に含まれる概念数，c_i は sd に含まれる i 番目の概念，そして $tf(\{d\}, c_i)$ は，ページの集合 $\{d\}$ に含まれるすべてのページにおける c_i の出現頻度の合計である．

4.5 ステップ 3：追加概念候補の検索とそれらの優先度付け

前述のステップ 2 で検索された Web ページ群に出現する語句が，基本的にはオントロジへの追加概念の候補である．しかし，検索されたそれぞれの Web ページには多数の不適切な文が含まれているので，それらをフィルタリングする．たとえば，コピーライトに関する記述や，著者紹介，宣伝広告を表す文が不適切な文に相当する．まず，検索された Web ページ中に出現する文の中で，サブドメインオントロジに含まれる概念を表す語句を含む文のみを残し，残りを除去する．そして，概念を表す語句と共起関係にある語句を追加概念の候補とする．候補となった語句の品詞に基づき追加する概念の型の候補を決める．次に，文中において，既存概念を表す語句と概念候補の語句の係り受け関係に着目し，候補をどの既存概念とどのような型の関連を用いて，サブドメインオントロジに追加するかの候補を決める．上記の概念の型および関連の型の候補選定は TCORE [9] での手法をそのまま利用している．具体的には，サ変動詞と動詞は function，形容詞は quality としている．名詞は object，actor，constraint のいずれかとして，オントロジ拡充者が選択することになる．既存概念と追加概念候補は Web ページ中の文において共起関係がある．既存概念と追加概念候補の型が，object と function の場合，apply 関係を候補とし，actor と function の場合，perform を候補とする．quality もしくは constraint が一方の型である場合，has-a 関係を候補とする．その他の場合はオントロジ拡充者が選択することになる．

我々は上記のフィルタリングによって不要な文を除去した Web ページ群を用いて，個々の既存概念ごとの視点から，概念候補を優先度付けする．我々は確率差（PD: probability difference）と term frequency × inverse document frequency（$TF \times IDF$）というメトリクスを用いて概念候補の語句を優先度付けする．PD はある文書に出現する 2 つの語句の共起関係に基づき定義されており，形態素解析ツール KH coder [30] で採用されている．$TF \times IDF$ によって，“the” や “http” 等の一般的過ぎる語句の優先度を下げる．$TF \times IDF$ は，語句の頻度とともに文書内での分布も考慮することで，当該語句が一般的か否かを予測している．

まず，PD の形式的な定義を以下に示す．複数の文から構成される文書 d に基づき，候補となる語句 t を既存概念 c に追加すべきか否かを検討しているとする．この設定において，PD は以下のように定義される．

$$PD(c,t,d) = \frac{lines(d,\{c,t\})}{lines(d,\{c\})} - \frac{tf(\{d\},t)}{line(d)}$$

上記において $line(d)$ は d に含まれる文の数，$lines(d,S)$ は d に含まれる文の中で，語句の集合 S に含まれるすべての語句が含まれる文の数，そして $tf(\{d\},t)$ は前述の NTF での補助定義と同様に，文書集合 $\{d\}$ に含まれる文書に出現する語句 t の数（頻度）である．t，c の共起頻度頻度が高ければ，PD は大きな値をとる．加えて，t，c の単独出現頻度が高ければ，たとえ共起頻度が高くても，PD が低い値をとるように調整されている．文書 d は既存概念 c が含まれる不要文書除去後の Web ページ群を連接したものである．

次に，$TF \times IDF$ の定義を以下に示す．$TF \times IDF$ は，ある語句 t が文書の集合 DS において頻繁に出現するが，DS に含まれる多数の文書には出現しないことを数値として以下のように表す．

$$TF{\times}IDF(t,DS) = tf(DS,t){\times}\left(\log\left(\frac{|DS|}{df(DS,t)}\right)+1\right)$$

上記において，$tf(DS,t)$ は，NTF や PD での補助定義と同様に，語句 t が文書の集合 DS に属する文書すべてに出現する数（頻度），$|DS|$ は DS に含まれる文書の数，$df(DS,t)$ は t が含まれる文書の数である．DS は概念候補 t が追加されるかもしれない既存概念を含む不要文書除去後の Web ページ群である．

4.6 ステップ 4：追加すべき概念の選択

このステップまででサブドメインに追加すべき新概念の候補がほぼ自動的に収集され優先度付けされている．この段階で実際に概念の追加を対話的に行う．ステップ 3 において，それぞれの既存概念に追加すべき概念候補が PD によって優先度付けされている．それぞれの候補の $TF \times IDF$ の値も知ることができる．これらの数値と概念候補となる語句自体を分析者が吟味し，既存概念に候補を新概念として追加する．

このような判断は分析者が自身の経験や従事しているプロジェクトの性質等を考慮して主観的に選択してよいが，手法の情報を以下のように活かすことを推奨している．まず，ある既存概念，たとえば「商品」に注目し，追加概念の候補群を閲覧する．閲覧に際しては，PD の大きい順に列挙することで，既存概念と共起しており，かつ，単独出現は少ないものが上位に配置される．ここで，候補として「IC タグ」という語句が上位にあったとする．この候補は，「商品に IC タグを付ける」や「IC タグがついた商品を…」等の文から候補となる．他の候補としては「登録する」，「削除する」等があるかもしれない．なお，既存概念との共起関係の多さにはかかわらず，出現頻度の多さと出現の局所性に注目したい場合は，PD の代わりに，$TF \times IDF$ を

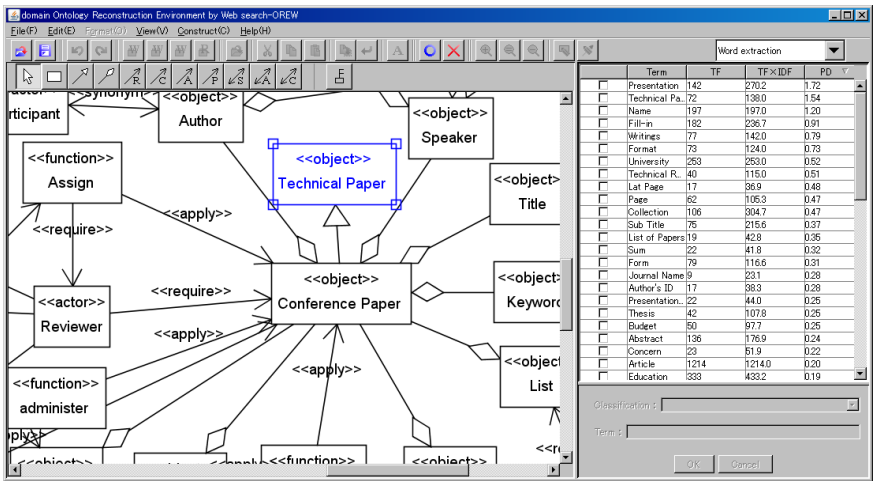


図 4 ツール OREW の画面例：候補の優先度付けと概念の追加

Fig. 4 A snapshot of OREW: Prioritizing candidates of concepts to be added.

用いて候補群を列挙してもよい．ステップ 3 で述べたように，既存概念と候補の品詞に基づき，手法は既存概念と候補間の関係案が提示されることもある．

5. OREW：オントロジ拡充法実施ツール

前章で述べたオントロジ拡充法を実施するため，我々は支援ツール OREW（domain Ontology Reconstruction Environment by Web search）を開発した．前章での拡充法を具体的に実施するためには，グラフ構造で表現された特定のオントロジの文法を想定しなければならない．現在の OREW の実装では，3 章で述べた ORE 法の文法を想定している．

OREW は前章でのステップ 1，2，3 の以下の設定を除いたすべてを自動的に遂行する．ステップ 1 では利用者は分割するサブドメインオントロジの数を指定する必要はある．ステップ 2 では OREW は “Yahoo! Web API” [*1] を利用して Web マイニングを自動的に行う．ステップ 3 では句読点に基づき検索された文書から文を自動的に認識するが，認識後の文の区切りを利用者が修正してもよい．また，追加概念の候補数を指定する必要はある．なお，Step 2 において収集される Web ページは，HTML およびテキストファイルに限られており，PDF，DOC，PPT 等の文書には対応していない．

ステップ 4 では，OREW は図 4 に示すような GUI を利用者に提供することで，利用者が新概念候補の選択と追加の支援を行う．図 4 の例では，既存概念 “Conference Paper” に新しい概念 “Technical Paper” が追加されたところを示している．サブドメイン数の数指定，目標ページ数の指定，概念候補の数の指定を除き，新しい概念 “Technical Paper” が追加される前の画面は，ツールですべて自動生成されている．ツール利用者は図に示すようにオントロジのグラフ構造をブラウズできる．図の右には “Conference Paper” への追加概念候補がリストとして列挙されている．それぞれの候補は TF，$TF \times IDF$，PD の数値が示されており，どれかの数値で整列させることができる．図では PD に基づき整列されている．4.6 節の後半で述べたように，$TF \times IDF$，PD の数値を参考に，どの候補から追加する検討を行うかを決める．なお，4.6 節の手法では言及されていないが，OREW では単純な語句の出現数である TF の情報も与えている．リストの左端のチェックボックスにチェックを入れることで，候補を実際に追加することができる．4.5 節で述べたように，追加の際には品詞に基づき概念の型の候補，および係り受け関係に基づき既存概念との関係の型の候補が示される場合がある．図 4 の例では，“Technical Paper” は名詞であるため，object，actor，constraint が該当するが，object を選択し，“Technical Paper” と “Conference Paper” は，名詞と名詞であるため，利用者が関係を設定しなければならない．画面例では is-a に設定している．なお，ツールの性能については 6.5 節で紹介する．

6. 評価

我々のオントロジ拡充法と，その支援ツール OREW を評価するために実験を行った．本章では実験の詳細と結果および考察を述べる．OREW はオントロジを拡充するツールであるため，本稿の評価点は拡充されたオントロジが拡充前のものよりも有効であるか否かである．オントロジ等のドメイン知識を用いた要求獲得が用いない場合よりも有効であることはすでに他の文献 [31] で発表されている．また，たとえドメイン知識が提供されたとしても，オントロジを利用するツールを用いた方が，用いない場合よりも有効であることもすでに他の文献 [10] で発表されている．よって，本章にあるようなオントロジの違いにのみ着目した実験を行った．

*1 http://developer.yahoo.com/

6.1 仮説

我々は拡充されたドメインオントロジが要求獲得結果の品質を向上させることを期待している．要求獲得結果の品質の中でも，完全性（Completeness）と正当性（Correctness）は重要な品質因子であり [32]，拡充されたオントロジが貢献可能な項目である．そこで，我々は以下の仮説にまず注目する．

H1：拡充されたオントロジで要求獲得を行うと，拡充前のオントロジで獲得を行うよりも，要求仕様書がより完全になる．すなわち，漏れの少ない要求仕様書となる．

H2：拡充されたオントロジで要求獲得を行うと，拡充前のオントロジで獲得を行うよりも，要求仕様書がより正当になる．すなわち，誤りの少ない要求仕様書となる．

もし，拡充されたドメインオントロジが獲得結果の向上に貢献するなら，拡充前後のオントロジの特徴の違いを見極めたい．我々のオントロジ拡充法および支援ツールでは，ドメインに特化した概念に，その補足となる概念が追加されてゆくように設計されている．また，我々は追加される概念は解決法 [33] 寄りの概念が多いと予測している．なぜなら，暗号化技術や IC タグ等，情報技術（IT）の解決法は異なる多数の分野（ドメイン）に利用可能だからである．以上をふまえ我々は以下の仮説を設定した．

H3：我々のオントロジ拡充法とツールによって追加される概念は，ドメインに特化したものというより，一般的なものである．

H4：我々のオントロジ拡充法とツールによって追加される概念は，対象世界の問題に関するものよりも，解決法に関するものの方が多い．

H5：追加される概念を，問題フレーム [33] で定義された causal，biddable，lexical の分類すると，なんらかの偏りがある．causal の概念は機械や電気部品等，因果律に大きく影響をうける概念である．biddable は人間のように自律的であり予測が難しい対象を表す概念である．lexical はデータや情報を表す概念である．

6.2 データと被験者

以下の 2 種類のオントロジを本実験の入力とした．これら拡充する前のオントロジを初期オントロジと呼ぶ．

- $\mathbf{IO_{pos}}$：POS（point of sales）システムのオントロジ．このオントロジはあるツールで [34] で生成されており，その生成には 14 個の POS システムに関する文書が利用された．生成に使われた文書のサイズはそれぞれ A4 用紙に 3 から 23 ページ程度である．
- $\mathbf{IO_{conf}}$：EasyChair 等の国際会議開催支援システムのオントロジ．このオントロジは国際会議運営の専門家によって作成された．作成に際して同専門家はいくつかの実システムの調査を行っていた．

IO_{pos} および IO_{conf} を用いた要求獲得におけるオントロジ参照の作業負担を均等にするため，概念の数がほぼ同じになるように調整を行った．オントロジによって獲得される要求項目の漏れのなさや誤りの少なさは，オントロジの規模に影響されない．しかし，獲得自体は人間の被験者が行うため，規模の違いによる被験者の労力の違いが，作業の集中力に影響を与える．この影響を排除するためオントロジの規模はほぼ同じとした．IO_{pos} はそのまま利用したが，IO_{conf} は is-a の下位階層，同義語，品質等の一部に相当する概念の一部を削除することで，IO_{pos} とほぼ同じ規模に調整した．実際の規模については 6.5 節で紹介する．

著者のうちの 1 人が OREW を用いて，IO_{pos} および IO_{conf} の拡充をそれぞれ行った．それぞれの拡充には，それぞれおよそ 4 時間を費やした．IO_{pos} を拡充したオントロジを $\mathbf{EO_{pos}}$，IO_{conf} を拡充したオントロジを $\mathbf{EO_{conf}}$ と呼ぶ．なお，EO_d（EO_d は EO_{pos} または EO_{conf}）は，IO_d に含まれるすべての概念と関連を含んでいる．

我々は 2 人の被験者 $\mathbf{Subject_a}$ および $\mathbf{Subject_b}$ にそれぞれのオントロジを用いて要求獲得を行うように依頼した．被験者はソフトウェア工学分野の学部 4 年生であり，プログラム言語やソフトウェア工学の基礎的知識は習得済みである．POS や国際会議開催支援業務に関する事前知識の差も被験者間にはないことを確認した．要求獲得には 3 章で説明した ORE 法とツールを利用してもらった．拡充前後のオントロジは被験者以外によって作成されているため，要求獲得は学習効果の影響を受けていないと考えられる．

3 章で説明した ORE 法とツールでは，初期要求のリストとドメインオントロジを入力として，要求獲得を行うことを想定している．そこで，我々は以下の 2 つの初期要求リスト（IRL: Initial Requirements List）を準備した．

- $\mathbf{IRL_{pos}}$：POS 分野の初期要求リスト．
- $\mathbf{IRL_{conf}}$：国際会議開催支援システムの初期要求リスト．

初期要求リストのサイズも 6.5 節で紹介する．3 章で説明した ORE 法とツールにおける要求獲得とは，初期要求リストをもとに，要求リストの項目数を増やすことである．もし，初期要求リストが拡充後のオントロジに有利なように準備されていた場合，実験の意味がないため，そうならないように初期要求リストを準備した．具体的には，図 1 における要求項目とオントロジ上の概念との対応付けに関しては，拡充前後のオントロジに差が出ないような初期要求リストを準備した．

6.3 実験設計

6.2 節で述べたように，著者の 1 人が IO_{pos} および IO_{conf} それぞれをもとに拡充を行い，EO_{pos} および EO_{conf} を得る．3 章で紹介した ORE 法とそのツールを被験者が学ぶ

表 1 それぞれの被験者の要求獲得の試行の順序
Table 1 The order of requirements elicitation for each subject.

	Subject_a	Subject_b
1 回目	IRL_stock and O_stock	
2 回目	IRL_conf and IO_conf	IRL_conf and EO_conf
3 回目	IRL_pos and EO_pos	IRL_pos and IO_pos

ために，別途，株価監視システムのオントロジ O_stock と初期要求 IRL_stock を準備し，被験者に手法とツールを学んでもらう．

前述の仮説 H1，H2 を確認するために，被験者に初期要求リストとオントロジを利用して要求獲得を実施してもらう．獲得の際，オントロジに加えて一般的な Web 検索を行うことも被験者に許可する．もし，一般的な Web 検索が頻繁に行われる場合，オントロジを用いた要求獲得法自体が支援となっていないことを示すことになる．この点を確認するために，あえて，一般的な Web 検索を禁止しなかった．学習効果による実験結果の影響を低減させるために，**表 1** にあるような順番で，それぞれの被験者に要求獲得を行ってもらう．要求獲得を行う回数が増えれば，被験者は獲得作業に慣れてくるため，結果は後に行われたものほど良くなる可能性がある．そこで，一方の被験者は拡張前のオントロジを用いた獲得の後に拡張後のオントロジを用いた獲得を行い，他方は逆順で獲得を行うことにする．具体的には，Subject_a には，まず IO_conf を用いた獲得を行ってもらい，次に EO_pos を用いた獲得を行ってもらう．反対に，Subject_b には，まず EO_conf を用いた獲得を行ってもらい，次に IO_pos を用いた獲得を行ってもらう．どちらの被験者も IO_d もしくは EO_d を用いた獲得作業を行う前に，獲得法を学ぶため，O_stock を用いた獲得を行ってもらう．獲得作業は 2 時間程度とするように依頼する．長時間の作業を行うと，作業に集中できなくなり結果にノイズが入る可能性がある．被験者が学生であることを考えて，授業時間より少し長い 2 時間を集中が持続可能な時間と判断した．

6.4 データ収集と測定

仮説 H1 と H2 のために，漏れなく誤りのない完全な要求項目リスト，すなわち正解の要求項目リストが必要である．我々は実験前に専門家の力を借りて，正解の要求項目リストを準備した．しかし，実験を通じて被験者は我々の見落とした正解となる要求項目を見つける場合もある．そこで，我々は正解の要求リストを以下のように定義する．

- 正解の要求リスト RRL_d は，事前に我々が準備した要求リストと，被験者が獲得した要求項目の中で正解と見なしてよいものの和集合である．なお，正解の要求リストには初期要求項目は含めないものとする．すなわち，$\text{IRL}_\text{d} \cap \text{RRL}_\text{d} = \phi$ である．

本実験では，ドメインオントロジを用いて要求獲得を行うことで，初期要求項目のリスト IRL_d が拡充される結果に注目している．以下に実験結果を評価するためのデータとなる拡充された要求項目リストの定義を示す．

- $\text{ERL}_\text{d}(\text{O}_\text{d})$：あるドメイン d の初期要求項目のリスト IRL_d をドメインオントロジ O_d を用いて拡充された要求項目リストである．なお，初期要求リストは拡充された要求リストから除外する．すなわち，$\text{IRL}_\text{d} \cap \text{ERL}_\text{d}(\text{O}_\text{d}) = \phi$ である．O_d は，実際には以下のどれかの値をとる：IO_pos，EO_pos，IO_conf，EO_conf．

$\text{ERL}_\text{d}(\text{O}_\text{d})$ は獲得を行う要求分析者（被験者）に依存するが，本実験ではその点を考慮しない．この点の考慮なしでも実験結果が影響を受けないように，最低限のソフトウェア工学の知識を持つ要求獲得の技量に大きく差のない者を被験者とした．

上記に定義した数値を用いて，以下のように，あるドメインオントロジの完全性（Completeness）および正当性（Correctness）を定義する．

- 完全性：$\text{Comp}(\text{O}_\text{d}) = \frac{|\text{RRL}_\text{d} \cap \text{ERL}_\text{d}(\text{O}_\text{d})|}{|\text{RRL}_\text{d}|}$
- 正当性：$\text{Corr}(\text{O}_\text{d}) = \frac{|\text{RRL}_\text{d} \cap \text{ERL}_\text{d}(\text{O}_\text{d})|}{|\text{ERL}_\text{d}(\text{O}_\text{d})|}$

これら 2 つは情報検索分野における recall と precision の指標にほぼ相当する．完全性と正当性は本来は要求項目 $\text{ERL}_\text{d}(\text{O}_\text{d})$ の性質ではある．しかし，この実験では，これらの性質は O_d によってもたらされた $\text{ERL}_\text{d}(\text{O}_\text{d})$ の性質といえるため，それぞれ $\text{Comp}(\text{O}_\text{d})$，$\text{Corr}(\text{O}_\text{d})$ のようにドメインオントロジ O_d の性質とした．実際には $\text{ERL}_\text{d}(\text{O}_\text{d})$ の完全性や正当性は獲得を行う分析者の力量や知識，作業時間等，他の要因も影響する．これらの影響を排除するために複数の被験者，一定の作業時間を設定した．なお，我々は被験者に対し，2 時間の作業時間が十分であったか否かを照会する．また，獲得中に一般的な Web サーチをどの程度利用したかも観察する．

仮説 H3，H4，H5 のため拡充前後のオントロジに含まれる概念に注目する．具体的には，オントロジの拡充によって，特定種類の概念の割合が増減したかに注目する．種類の詳細については後述する．このような増減を測定するために，我々は以下のような関数を導入し，概念の数や割合を測定した．

- Concept(X)：あるオントロジ X に含まれる概念の集合．
- Type(C, d, t)：概念の集合 C に含まれる t 型の概念をすべて集めた概念の集合．C に含まれる概念はドメイン d のドメインオントロジに含まれる．ある概念がある種類 t 型であるか否かは，場合によっては分析者が主観的に判断しなければならない．よって，この関数の計算には分析者の主観が関わる．
- 集合演算に一般的に用いられる演算子も用いる，たと

えば集合 S の要素数 $|S|$，補集合 $\bar{S}$，差集合 $S-T$ 等．

上記に基づきそれぞれのドメイン d について下記の値を導出する．

- $\mathrm{Concept(IO_d)}$：ドメイン d の初期オントロジに含まれる概念集合．
- $\mathrm{Concept(EO_d)} - \mathrm{Concept(IO_d)}$：オントロジ拡充によって追加された概念の集合．
- $\frac{|\mathrm{Type(Concept(EO_d)-Concept(IO_d),d,t)}|}{|\mathrm{Type(Concept(IO_d),d,t)}|}$：t 型の概念の増加率．これを我々は $\mathbf{Gain(d,t)}$ と略記する．その理由は電気等の増幅率（ゲイン）と意味が似ているからである．

この実験では以下の 4 つの概念に関する分類に着目する．それぞれの分類において，相互の型は排他的である．

- general 型もしくは specific 型：
  ある概念があるドメインに特化した概念である場合，specific 型であり，そうでなければ general 型に分類する．以下の例では，”paper submission”は会議運営 (”conference management”) 分野に特化した概念だが，”password”はそうではないため，前者は specific 型，後者は general 型となっている．
  Type({”paper submission”,”password”},
  ”conference management”,”general”)
  = {”password”}
- solution 型もしくは problem 型：
  ある概念が暗号化や IC タグ等の解法に関することの場合，その概念を solution 型と見なし，それ以外を problem 型と見なす．
- causal 型，biddable 型もしくは lexical 型：
  これらの型は問題フレーム [33] に由来する．ある概念が機械や電気部品のように因果律に従う場合，その概念を causal とする．ある概念が人間のように自律的で予測が困難なものである場合，その概念を biddable 型とする．ある概念がデータや情報を示す場合，その概念は lexical 型とする．
- any 型：
  すべての概念は any 型を持つと定義する．この型は単に型に関係なくオントロジ拡充の増幅率（ゲイン）を知るために用いる．

6.5 結果

最初に**表 2** に初期および拡充されたオントロジの規模を示す．前述のとおり拡充は著者の 1 人によって OREW を用いて行われた．それぞれのドメインにおいて，拡充は 4 時間ほどの時間を要した．OREW の運用に際して，サブドメイン数は 7 とし，目標ページ数は 20，概念候補数は 20 と設定した．1 つのサブドメインを構成する概念が平均して 3 個程度になるように，7 という数値は設定された．入力とするオントロジの規模が 40 概念程度であり，運用経験より，およそ半数が名詞となるためである（40/7/2=2.8）．目標ページ数は 3 個の検索語の AND 検索の試行より決めた．概念候補数の 20 個も経験的に，20 個の候補を設定すれば，漏れは少なく，また，図 4 に示した UI においても一覧性があるため，この数値が設定された．4 時間の拡充時間のうち，後半の 2 時間が，図 3 の Step 4 に費やされた．表 2 にあるように初期概念数が 40 個程度なので，1 つの既存概念に対する追加概念候補の検討には平均して 3 分ほど費やしている．よって，人間が最終的に候補を採択する作業が大きな負担ではないと考えられる．実際，拡充を行った者に，この点を照会したが，利用性に関しての苦情はなかった．表 2 に示すように，結果として追加された概念は既存の 1 概念あたり 1 個程度である．追加された概念がなかった場合から 4 個の場合もあり，既存概念によってばらつきは大きい．候補が 20 個提示されるため，候補の採択率は 5%程度（1/20）となり低い値である．たとえば，ドメイン conf における「論文」への追加候補として「教育」,「免許」,「科学研究費」があがったが，これらは採択されなかった．20 個すべて採択されることはなかったため，本来候補となるべき概念が候補から漏れていたことはないと考えられる．

表 2 オントロジの規模（注：練習で使ったオントロジの規模は 69）
Table 2 The sizes of ontologies (Note: $|\mathrm{Concept(O_{stock})}|$ is 69).

d	pos	conf
$\lvert\mathrm{Concept(IO_d)}\rvert$	43	42
$\lvert\mathrm{Concept(EO_d)}\rvert$	94	98
$\lvert\mathrm{Concept(EO_d)} - \mathrm{Concept(IO_d)}\rvert$	51	56
Gain(d,”any”)	118.6%	133.3%

Step 1，3 はほとんど時間がかかっておらず，Step 2 がおよそ 2 時間かかっている．ページの検索およびダウンロード自体，時間がかかることと，目標ページ数（この実験では 20 ページ）に達しない場合，サブドメインから概念を減らして，再試行を行うため，このように長い時間がかかった．再試行も自動的に行われるため，拡充を行う人間が関与する必要はない．なお，20 ページ程度のページを収集すれば，適切な追加概念の候補を含むページを収集できることを，ツールの動作テストを通して調べた．本実験では，Core2duo 2 GHz，メモリ 2 GB，Windows XP OS のラップトップを用いて行った．通信回線は有線のイーサネット（100Base-T）である．

表 2 に示すようにどちらのオントロジもおよそ 2 倍の規模となっているため，提案手法および OREW によるオントロジの増幅率（ゲイン）はおよそ 100%となる．なお，拡充されたオントロジについては，規模を揃えるための概念の削除等は行っていない．

次に**表 3** に仮説 H1，H2 に関係するデータである初期および拡充された要求リスト，およびそれらの完全性およ

表 3 要求リストと拡充と完全性，正当性
Table 3 Requirements lists and their Completeness and Correctness.

d	pos		conf	
O_d	IO_{pos}	EO_{pos}	IO_{conf}	EO_{conf}
$\|IRL_d\|$	11		10	
$\|ERL_d(O_d)\|$	26	58	33	65
$\|RRL_d\|$	71		91	
$\|ERL_d(O_d) \cap RRL_d\|$	18	50	29	59
$Comp(O_d)$	24.4%	70.4%	31.9%	64.8%
$Corr(O_d)$	69.2%	86.2%	87.9%	90.8%
2 時間で十分か？	Yes	No	Yes	No

び正当性を示す．表の読み方を，POS ドメインについて拡充前のオントロジ IO_{pos} を用いた獲得を利用して説明する．POS ドメインの初期要求リスト IRL_{pos} は 11 項目の要求文を含んでいた．それが，拡充前のオントロジ IO_{pos} によって，26 個の項目が増えた．正解の要求項目は 71 項目だが，拡充された 26 個のうち，18 個が正解に含まれていた．よって，完全性および正当性はそれぞれ 24%，69%となった．結果として，どちらのドメインの場合も，完全性および正当性の双方が改善されている．

表 3 の最終行は 2 時間の獲得時間が十分か否かを被験者に尋ねた結果を示している．拡充前のオントロジ（IO_d）を用いた場合，2 時間は十分であると答えたが，拡充後のオントロジ（IE_d）を用いた場合は不足していると回答した．しかし，被験者らには事前に 2 時間程度で終えるように依頼したため，拡充後のオントロジ（IE_d）を用いた場合でも 2 時間に収まるように作業をしたという回答も得た．実験中の被験者の観察から，一般的な Web 検索はほとんど行われていなかったことが分かった．この結果より，オントロジを用いた要求獲得法それ自体が支援となっていたことが分かった．

最後に**表 4**，**表 5** および **表 6** に，オントロジ拡充によって生じた，それぞれの分類における概念の型ごとの増幅率を示す．たとえば，表 4 では，Gain(d,"general") が 307.6%であり，これは，POS のドメインオントロジを OREW で拡充したところ，general 型の概念に関して，もとからある概念に，およそ 3 倍の概念が追加されたということを示している．

6.6 議論

仮説 H1 と H2 では，拡充されたオントロジが獲得される要求項目の品質に与える影響に注目している．表 3 に示すように，要求項目の完全性は，拡充されたオントロジによって，POS および国際会議開催支援どちらの場合も 2 倍程度に改善されている．よって，仮説 H1 は正しいと考えられる．オントロジ拡充によって情報が増えているため，完全性が改善されることは，容易に予測される結果である．

表 4 general 型もしくは specific 型概念の増幅率
Table 4 Increasing Rate of general or specific concepts.

d	pos		conf	
Gain(d,"general")	307.6%	(40/13)	163.1%	(31/19)
Gain(d,"specific")	36.6%	(11/30)	108.6%	(25/23)

表 5 problem 型もしくは solution 型の増幅率
Table 5 Increasing Rate of problem or solution concepts.

d	pos		conf	
Gain(d,"solution")	133.3%	(20/15)	212.5%	(17/8)
Gain(d,"problem")	110.7%	(31/28)	114.7%	(39/34)

表 6 Icausal 型，biddable 型もしくは lexical 型の増幅率
Table 6 Increasing Rate of causal, biddable or lexical concepts.

d	pos		conf	
Gain(d,"causal")	150.0%	(12/8)	350.0%	(21/6)
Gain(d,"biddable")	215.3%	(28/13)	100.0%	(17/17)
Gain(d,"lexical")	50.0%	(11/22)	94.7%	(18/19)

要求項目の正当性についても，微増ではあるが，拡充されたオントロジによって改善されている．よって，仮説 H2 も正しいと考えられる．一般に提供される情報が単に増加するだけでは，正確さに相当する正当性は誤った情報や不要な情報によって減少する場合が多い．しかし，本実験からは，完全性だけでなく，正当性も改善されており，要求獲得結果の漏れのなさ（完全性）および誤りの少なさ（正当性）の双方に貢献しているといえる．これは，提案手法によるオントロジの拡充は，ドメインに特化した概念をより正確に理解することに貢献した結果であると思われる．

仮説 H3，H4，H5 では，我々はどのような種類の概念が提案したオントロジ拡充法で増幅されるかに注目している．そこで，表 2，4，5 および 6 の Gain（増幅率）の値を考察する．表 2 に示すように，概念の種類を区別しなければ，増幅率はおよそ 100%であるが，他の表にある種類別の増幅率は 100%とは異なる値のものが多い．表 4 に示すように，POS および会議支援分野のどちらの場合も，general 型の概念（一般的な概念）は，specific 型の概念に比べ増幅率が大きいため，仮説 H3 は正しいと思われる．同様に，表 5 の結果から H4 も正しいと思われる．表 6 の結果から，POS および会議支援双方において，causal 型の概念が lexical 型の概念よりも増幅率が大きい．仮説 H5 は仮説というより模索的な質問に相当する．よって，この仮説に関しては，我々のオントロジ拡充法は causal 型の概念の追加を促進していると述べることができる．causal 型の概念は，機械や電気部品等の外部環境に相当するため，特定分野には特化はしていないが，業務を理解するうえで提供するに値する概念と思われる．よって，我々のオントロジ拡充法が，当初の目標どおりに機能していることの一端を H5 から確認することができた．

6.7 妥当性への脅威

ソフトウェア工学の評価実験において，実験方法と結果の妥当性を脅かす因子を明確にすることが求められている．たとえば，本実験において単一被験者が，最初に拡充前のオントロジを使った要求獲得，次に拡充後のオントロジを使った要求獲得を順に行った場合，学習効果によって，後者に有利なバイアスがかかっている疑いが生じる．近年，このような因子は “Threat to Validity” という見出しを用いて，以下の 4 つの視点から行うことが一般的になりつつある [35], [36]．本稿での実験も，この 4 つの視点から，実験方法と結果が妥当であるか否かについての考察を行う．

6.7.1 内部妥当性

実験データ（独立変数）に悪影響を与える要因の有無を検討する．本実験での独立変数はオントロジと要求項目リストである．要求項目リストについては，表 1 に示すように，同じ被験者が同じリストを 2 回利用しないようにすることで，学習効果の影響を排除した．初期要求リストに含まれる概念群は拡充前のオントロジに含まれるものがほとんどである．よって，拡充されたオントロジに有利なようには設計されていない．オントロジの拡充は著者の 1 人が逐次的に行ったため，学習効果は排除されていない．しかし，本実験では拡充された 2 つのオントロジ EO_{pos} と EO_{conf} の比較を行っているわけではないので，この点は内部妥当性への脅威とならない．6.2 節に述べたように，実験前に，IO_{conf} は IO_{pos} とサイズを揃えるため，概念の削除を行っているため，拡充されたオントロジに有利なようにバイアスがかかっている恐れはある．しかし，提案したオントロジ拡充法によって拡充が期待される概念を優先的に削除したわけではないため，バイアスは小さい，もしくはないと考える．

6.7.2 外部妥当性

実験結果が一般的か否かを検討する．本実験ではオントロジの拡充をした被験者と，それを用いて要求獲得を行った被験者は異なる．よって，この点における学習効果による影響はないという意味で一般的である．提案したオントロジ拡充法の一部は拡充を行う者の主観に依存している．また，本実験ではたった 1 人の被験者が拡充を行った．よって，他の者が拡充を行った場合，異なるオントロジに拡充される可能性があるという点において，一般性が低い．拡充および獲得を行った被験者は学生であるため，一般の技術者が行った場合とは異なる結果になっている可能性があるという点で一般性が低い．

6.7.3 構成妥当性

測定したデータが測定したいことを反映しているか否かを検討する．我々は仮説群で利用する変数を直接的に測定している．しかし，H3，H4，H5 では概念については測定を行っているが，概念間の関係については測定を行っていない．また，ある概念が一般的であるか特殊であるか等の型分類は，一部的，実験者の主観に頼る部分があるため，構成妥当性への脅威となる．仮説 H1，H2 については正しい要求項目リスト RRL_d が必要であった．RRL_d と拡充されたオントロジの双方は著者らによって作成されたため，これら 2 つに因果関係があるかもしれない．表 3 の最終行に示すように，拡充されたオントロジを用いた場合，2 時間以上の作業時間が必要であったように見られる．作業時間を 2 時間に制限しなかった場合，H1，H2 に関係する完全性，妥当性の数値は変化した可能性がある．しかし，時間不足を申し出たのは，拡充したオントロジ側であり，時間の制限をなくしたり，休憩を入れたうえで作業を継続したとしても，H1，H2 の仮説に有利な変化となると思われる．

なお，本実験では，オントロジの増幅（H3，H4，H5）が原因となり，要求獲得結果の改善（H1，H2）となったことの分析は行っていない．よって，要求獲得結果の改善（H1，H2）とオントロジの増幅（H3，H4，H5）の双方に影響を与える因子が存在することは否定できない．

6.7.4 結論妥当性

同様の実験を行った場合，同じ結果が得られるか否かの脅威を検討する．データが 2 件のみであるため，統計的検定を行っておらず，妥当性への脅威が残る．

7. 結論

本稿では要求獲得のためのドメインオントロジを拡充するための手法を提案し，その評価を行った．ドメインオントロジがドメインに特化した文書や専門家のみから抽出したものである場合，そのドメインオントロジを利用した要求獲得において，要求の抜けや誤りが生じる場合がある．提案手法の目標は，このようなドメインに特化したオントロジを要求分析者が理解するのに有益な概念を追加することである．そのような追加概念候補を探すために，我々は Web マイニングと軽量化自然言語処理の技術を利用した．我々は手法を実施するための支援ツール OREW を開発し，拡充されたオントロジが有用なものとなっているか否かの評価を行った．実験を通した評価の結果，我々の手法で拡充されたオントロジは，要求獲得結果の品質向上に貢献していることが確認できた．

4 章に述べたように，拡充法の最終ステップ以外は自動的に遂行できる．しかし，追加概念候補を選択する部分は分析者の主観的な判断を必要とする．今後は特に追加概念候補の型および既存概念との関係の予測部分を改善したい．現状では一般的な電子辞書やオントロジ等（たとえば，WordNet [37]）を利用していない．よって，Web ページに出現する同義語等の処理を自動的に行うことができない．この部分についても，今後，一般的な電子辞書やオントロジ等を利用することで改善したい．

参考文献

[1] Breitman, K.K. and do Prado Leite, J.C.S.: Ontology as a Requirements Engineering Product, *11th IEEE International Requirements Engineering Conference* (*RE'03*), Monterey Bay, California, USA, pp.309–319 (2003). Mini-Tutorial.

[2] Lee, S.W. and Gandhi, R.A.: Ontology-based Active Requirements Engineering Framework, *APSEC*, pp.481–490 (2005).

[3] Kaiya, H. and Saeki, M.: Using Domain Ontology as Domain Knowledge for Requirements Elicitation, *Proc. 14th IEEE International Requirements Engineering Conference* (*RE'06*), Minneapolis/St. Paul, Minnesota, USA, IEEE CS, pp.189–198 (2006).

[4] Dzung, D.V. and Ohnishi, A.: Improvement of Quality of Software Requirements with Requirements Ontology, *International Conference on Quality Software*, pp.284–289 (online), DOI: http://doi.ieeecomputersociety.org/10.1109/QSIC.2009.44 (2009).

[5] Pohl, K., Bockle, G. and Linden, F.V.D.: *Software Product Line Engineering: Foundations, Principles And Techniques*, Springer-Verlag New York Inc. (2005).

[6] Gruber, T.R.: A translation approach to portable ontologies, *Knowledge Acquisition*, Vol.5, No.2, pp.199–220 (1993).

[7] Zhao, Y., Dong, J. and Peng, T.: Ontology Classification for Semantic-Web-Based Software Engineering, *IEEE Trans. Services Computing*, Vol.2, pp.303–317 (online), DOI: http://doi.ieeecomputersociety.org/10.1109/TSC.2009.20 (2009).

[8] Zhou, H., Chen, F. and Yang, H.: Developing Application Specific Ontology for Program Comprehension by Combining Domain Ontology with Code Ontology, *QSIC*, pp.225–234 (2008).

[9] Kitamura, M., Hasegawa, R., Kaiya, H. and Saeki, M.: An Integrated Tool For Supporting Ontology Driven Requirements Elicitation, *ICSOFT 2007, 2nd International Conference on Software and Data Technologies*, Barcelona, Spain, pp.73–80 (2007).

[10] Kitamura, M., Hasegawa, R., Kaiya, H. and Saeki, M.: A Supporting Tool for Requirements Elicitation Using a Domain Ontology, *Software and Data Technologies*, Vol.22, pp.128–140, Springer Berlin Heidelberg (online), DOI: 10.1007/978-3-540-88655-6 (2008). Communications in Computer and Information Science (CCIS).

[11] Kato, J., Komiya, S., Saeki, M., Ohnishi, A., Nagata, M., Yamamoto, S. and Horai, H.: A Model for Navigating Interview Processes in Requirements Elicitation, *APSEC*, pp.141–148 (2001).

[12] Capobianco, G., Lucia, A.D., Oliveto, R., Panichella, A. and Panichella, S.: On the role of the nouns in IR-based traceability recovery, *ICPC*, pp.148–157 (2009).

[13] Jurafsky, D. and Martin, J.: *Speech and Language Processing*, Prentice Hall (2000).

[14] Keenan, E.L.: *Formal Semantics of Natural Language*, Cambridge University Press (1975).

[15] Dong, J.S., Feng, Y., Li, Y.-F. and Sun, J.: A Tools Environment for Developing and Reasoning about Ontologies, *APSEC*, pp.465–472 (2005).

[16] Zouaq, A. and Nkambou, R.: Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project, *IEEE Trans. Knowl. Data Eng.*, Vol.21, No.11, pp.1559–1572 (2009).

[17] Li, M. and Zang, F.: A Self-Feedback Methodology of Domain Ontology Modeling, *World Congress on Software Engineering*, Vol.2, pp.218–223 (online), DOI: http://doi.ieeecomputersociety.org/10.1109/WCSE.2009.178 (2009).

[18] Wang, X., Chen, P., Wang, X. and Liu, P.: Research on Chinese Domain Ontology Modeling Based on Automatic Knowledge Acquirement from Multiple Dictionaries, *International Symposium on Knowledge Acquisition and Modeling*, pp.360–366 (online), DOI: http://doi.ieeecomputersociety.org/10.1109/KAM.2009.216 (2009).

[19] Luong, H.P., Gauch, S. and Wang, Q.: Ontology Learning Through Focused Crawling and Information Extraction, *International Conference on Knowledge and Systems Engineering*, pp.106–112 (online), DOI: http://doi.ieeecomputersociety.org/10.1109/KSE.2009.28 (2009).

[20] Storey, V.C., Chiang, R.H.L. and Chen, G.L.: Ontology Creation: Extraction of Domain Knowledge from Web Documents, *ER*, pp.256–269 (2005).

[21] Nakayama, K., Hara, T. and Nishio, S.: Wikipedia Mining for an Association Web Thesaurus Construction, *WISE*, pp.322–334 (2007).

[22] Durham, J., McLauchlan, L. and Yuster, R.: Enabling a Common and Consistent Enterprise-Wide Terminology: An Initial Assessment of Available Tools, *Web Intelligence*, pp.544–548 (2008).

[23] Mikroyannidis, A. and Theodoulidis, B.: Heraclitus II: A Framework for Ontology Management and Evolution, *Web Intelligence*, pp.514–521 (2006).

[24] Subramaniam, L.V., Nanavati, A.A. and Mukherjea, S.: Enriching One Taxonomy Using Another, *IEEE Trans. Knowl. Data Eng.*, Vol.22, No.10, pp.1415–1427 (2010).

[25] Kong, H., Hwang, M. and Kim, P.: A New Methodology for Merging the Heterogeneous Domain Ontologies Based on the WordNet, *International Conference on Next Generation Web Services Practices*, pp.235–240 (online), DOI: http://doi.ieeecomputersociety.org/10.1109/NWESP.2005.7 (2005).

[26] Maree, M. and Belkhatir, M.: A Coupled Statistical/Semantic Framework for Merging Heterogeneous Domain-Specific Ontologies, *IEEE International Conference on Tools with Artificial Intelligence*, Vol.2, pp.159–166 (online), DOI: http://doi.ieeecomputersociety.org/10.1109/ICTAI.2010.138 (2010).

[27] Tanabe, D., Uno, K., Akemine, K., Yoshikawa, T., Kaiya, H. and Saeki, M.: Supporting Requirements Change Management in Goal Oriented Analysis, *Proc. 16th IEEE International Requirements Engineering Conference* (*RE'08*), Barcelona, Catalunya, Spain, IEEE CS, pp.3–12 (2008).

[28] Goh, K.-I., Oh, E., Kahng, B. and Kim, D.: Betweenness centrality correlation in social networks, *Phys. Rev. E*, Vol.67, No.1, p.017101 (online), DOI: 10.1103/PhysRevE.67.017101 (2003).

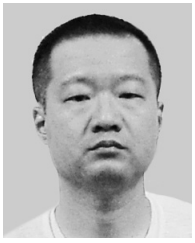
[29] Fallaw, W.C.: A Test of the Simpson Coefficient and Other Binary Coefficients of Faunal Similarity, *Journal of Paleontology*, Vol.53, No.4, pp.1029–1034 (1979).

[30] KH Coder, available from ⟨http://sourceforge.net/projects/khc/⟩ (accessed 2011-01).

[31] 加藤潤三，佐伯元司，大西　淳，海谷治彦，山本修一郎：シソーラスを利用した要求獲得方法（THEOREE），情報処理学会論文誌，Vol.50, No.12, pp.3001–3017 (2009).

[32] IEEE Recommended Practice for Software Requirements Specifications, IEEE Std. 830-1998 (1998).

[33] Jackson, M.: *Problem Frames, Analyzing and structuring software development problems*, Addison-Wesley (2000).
[34] Hasegawa, R., Kitamura, M., Kaiya, H. and Saeki, M.: Extracting Conceptual Graphs from Japanese Documents for Software Requirements Modeling, *Proc. 6th Asia-Pacific Conference on Conceptual Modelling* (*APCCM 2009*), Wellington, New Zealand, pp.87–96 (2009). Vol.96 in the Conferences in Research and Practice in Information Technology Series.
[35] Wohlin, C., Runeson, P., Host, M., Ohlsson, M.C., Regnell, B. and Wesslen, A.: *Experimentation in Software Engineering An Introduction*, Kluwer (2000).
[36] Kitchenham, B., Pfleeger, S.L., Pickard, L., Jones, P., Hoaglin, D.C., Emam, K.E. and Rosenberg, J.: Preliminary Guidelines for Empirical Research in Software Engineering, *IEEE Trans. Softw. Eng.*, Vol.28, No.8, pp.721–734 (2002).
[37] Miller, G.A.: WordNet: A Lexical Database for English, *Comm. ACM*, Vol.38, No.11, pp.39–41 (1995).



海谷 治彦 （正会員）

1994 年東京工業大学博士（工学）取得．現在，信州大学工学部准教授，国立情報学研究所客員准教授．

清水 悠太郎

2010 年信州大学学士取得．

安井 浩貴

2010 年信州大学修士（工学）取得．

海尻 賢二 （正会員）

1977 年大阪大学工学博士取得．現在，信州大学工学部教授．

林 晋平 （正会員）

2008 年東京工業大学博士（工学）取得．現在，東京工業大学助教．

佐伯 元司 （正会員）

1983 年東京工業大学工学博士取得．現在，東京工業大学教授．国立情報学研究所客員教授．