

記述形式の多様性を考慮した 電子メール中の議論過程抽出手法

中山 祐貴^{1,a)} 大沼 亮^{1,b)} 神長 裕明^{1,c)} 横山 節雄^{2,d)} 宮寺 庸造^{2,e)} 中村 勝一^{1,f)}

受付日 2014年4月14日, 採録日 2014年10月8日

概要: 研究活動をはじめとする様々な知的生産活動の中で、電子メールを用いた議論を行う機会が多い。議論の過程・結果の把握は重要だが、扱う電子メール数の多さや増加速度ゆえに、一般にユーザにとっては容易な作業ではない。これまでに議論過程理解支援に関する研究がいくつか報告されているが、対象が限定的なものがほとんどで、実効的な支援は得られていない。特に、実際の電子メールにおける多様な記述スタイルに対応可能な手法は実現されていない。そこで本研究では、日々蓄積していく電子メールファイル塊中から、できるだけ強い制約を課さず、記述形式の多様性に対応しうる方法で議論過程を抽出し、ユーザに視覚的に提示する新たなアプローチによる議論過程把握支援の実現を目指している。本論文では主に、記述形式の多様性を考慮した議論過程抽出手法について述べる。また、実際の電子メールデータを用いた実験に基づいて、提案手法の有効性や特徴について考察する。

キーワード: 議論過程, 話題抽出, 応答ペア, 電子メール, 議論支援

Methods for Extraction of Discussion Processes in E-mails Considering the Diversity of Description Style

HIROKI NAKAYAMA^{1,a)} RYO ONUMA^{1,b)} HIROAKI KAMINAGA^{1,c)}
SETSUO YOKOYAMA^{2,d)} YOUZOU MIYADERA^{2,e)} SHOICHI NAKAMURA^{1,f)}

Received: April 14, 2014, Accepted: October 8, 2014

Abstract: Discussion using e-mail is frequently conducted in intelligent works such as research activity. It is important but generally difficult to successfully grasp the processes of the discussions and their results. Although there have been some research projects which aim to support grasping of discussion processes, these existing methods have not effective enough because of unreasonable restrictions. In particular, most existing methods have not considered the actual e-mail features such as diversity of description style. Therefore, this research project has aimed to develop methods for extraction of discussion processes and their visualization without hard restrictions. This paper mainly describes methods for extracting the discussion processes in e-mails considering the diversity of description style. Finally, effectiveness and characteristics of the methods are discussed based the results of the experiments using practical e-mail data.

Keywords: discussion process, topic extraction, response pair, e-mail, discussion support

¹ 福島大学大学院共生システム理工学研究科
Graduate School of Symbiotic System Science and Technol-
ogy, Fukushima University, Fukushima 960–1296, Japan

² 東京学芸大学教育学部
Faculty of Education, Tokyo Gakugei University, Koganei,
Tokyo 184–8501, Japan

a) nakayama@cs.sss.fukushima-u.ac.jp

b) onuma@cs.sss.fukushima-u.ac.jp

c) kami@sss.fukushima-u.ac.jp

d) yokoyama@u-gakugei.ac.jp

1. まえがき

企業や研究教育機関における知的生産活動の中で、プロジェクトメンバなどと議論を行う手段として電子メールを用いる機会が多々ある。ここで述べる議論とは、ある目

e) miyadera@u-gakugei.ac.jp

f) nakamura@sss.fukushima-u.ac.jp

的・話題について関係者同士で発言し合うことを指す。また、発言の積み重ねによる議論の経過や結果を議論過程と呼ぶ。議論の円滑な進行、議論で得られた成果・知見の活用のためには、議論過程をうまく把握することが重要だが、容易ではない。その主な要因として、まず、ユーザのメールアプリケーションには、通常大量のメールが蓄積されていることがあげられる。その内訳も、過去に議論した電子メール、今現在議論している電子メール、ダイレクトメールのような議論とはまったく関係のないメールなど多様である。そのため、これら大量のメールの中から、ユーザが注目する議論に関するメールを抽出するには相当の手間を要する。

これに加えて、一般にユーザはつねに1つの議論だけを行っているとは限らず、複数の議論を並行して処理することが多い。そのため、複数の話題を把握し、それぞれの議論に的確に発言しなければならない。ゆえに、議論に関する電子メールを話題ごとに整理し、さらにその流れを把握することが必要となるが、実際には負担が大きく難しい作業である。

これらの問題に対して、メール中の話題の把握支援を目指したシステム [1] や議論の流れの視覚化 [2] などの研究が報告されているが、対象とするメールに関する強い制約などのため、電子メールを使った実際の議論における議論過程把握について十分に有効な支援は実現されていない。

そこで本研究では、強い制約を課さずに、できるだけ通常使用の電子メールからそのまま議論過程を抽出し、ユーザに視覚的に提示する新たなアプローチによる議論過程把握支援の実現を目指す [3]。

これにより、ユーザによる議論過程の容易な把握を可能とし、議論自体への専念・議論の円滑化が期待できる。本論文では主に、記述形式の多様性を考慮した議論過程抽出手法について述べる。

本論文では、2章で、電子メールを用いた議論の問題点と本研究の方針を述べる。3章で議論過程抽出手法について詳述する。また、4章で、提案手法に基づいた議論過程把握支援システムの概略を示す。さらに、5章で実際の電子メールデータを用いた実験について報告し、その結果に基づいて提案手法の有効性について考察する。最後に6章で、本論文のまとめと今後の課題について述べる。

2. 問題点と支援方針

2.1 電子メールを用いた議論の問題点

まず、電子メールは議論専用の媒体ではなく、一般的なコミュニケーション媒体である。ゆえに、個々のユーザが保持するメール群には、議論に関係しないメールも含まれている。また、メール本文中には、議論とは関係のない文章、たとえば挨拶文のような定型文も記述されている。さらに、複雑な議論の場合は、1つの電子メール中に複数の議

論に関する記述が含まれることもある。そのため、特定の議論のみに使用する専用メールアドレスを準備している場合や、議論のために専用の記述フォームなどのメタデータ構造を持つ他の媒体に比べると、議論に関する記述の抽出は難しい。つまり、電子メール群から議論に関するメール・記述箇所を取り出すことは容易ではない(問題点1)。

次に、あるメールに対して返信する際に、返答の対象となる記述を明示する規則は、日常のメール使用では存在しない。本文中に返答先の記述を貼り付けて引用する記述方法が比較的なじみ深いですが、その方法は個々のユーザの習慣などによりまちまちで、引用がなされないことも多い。そのため、議論が長期間に及ぶ場合や議論に関わる人数が多い場合には、議論に関するメールと記述数が膨大となり、メールや記述どうしのつながりも複雑になる。さらに、複数の議論を並行して行っている場合は、把握すべき議論の数が増加する。ゆえに、電子メール中の記述どうしのつながりや複雑な議論の流れを把握することが難しくなる。つまり、議論の円滑な進行などのために重要であるにもかかわらず、ユーザは議論過程を把握することが難しい(問題点2)。

2.2 記述粒度と応答記述形式の多様性

電子メール中の議論に関する記述とその間のつながりの抽出を妨げている問題について整理する。まず、電子メール本文中で議論に関係する部分を考えてみると、その記述の範囲は一定ではない。メール本文中の1文章などごく短い記述だけが議論に関わる場合や、本文中の記述すべてが議論に関するものである場合など様々である。

次に、議論の流れを抽出するためには、メール記述間の返答関係を解析する必要がある。ここで、あるメールに対して返信する場合に、返答対象となる文章を引用して隣接行に返答を記述する行為(応答記述と呼ぶ)に着目する。通常の電子メールでは、この応答記述の形式は多様であり、その記述粒度によって、いくつかの記述パターンに分かれる(表1)。

電子メールを用いた実際の議論では、これら記述粒度の混在は一般的であるため、2.1節で指摘した「議論に関するメール・記述箇所の抽出」、「電子メール中の記述どうしのつながりの抽出」をいっそう難しくしている。それゆえに、電子メール中の議論に関する記述やその間のつながりの抽出を目指した既存研究は、対象メールの制限や記述方法に関する強い制約を課さざるをえなかった。このことが、通常の電子メールからそのまま議論過程を抽出しうる手法がこれまで実現されてこなかった1つの大きな要因と考えられる。したがって、強い制約を課さずに通常の電子メールから議論過程を取り出すために、これら記述粒度・応答記述形式の多様性に対応しうる方法が必要である。

表 1 応答の形式と記述例

Table 1 Style of reply and examples of their descriptions.

応答の形式	記述例
1 文章 → 1 文章	>引用文
	返答文
1 文章 → 複数文章	>引用文 >引用文
	返答文
電子メール → 1 文章	>引用文
	[返答メール全文] 返答文 返答文
電子メール → 電子メール	[返答メール全文] 返答文 返答文
	[返答先のメール全文] >引用文 >引用文

2.3 関連研究

協調学習などのフィールドでの議論支援 [4], [5] が報告されている。特定活動に的を絞り、そこでの議論の円滑化や多角的展開を支援しており興味深い。しかし、システムによる議論状態の推定や停滞議論について発言を促す助言提示などを主眼としており、本研究とは立場・対象が異なる。

電子メールによる会話の要約に関する研究 (たとえば文献 [1], [6]) がさかんに行われている。すでに終了した議論について結果に主眼をおいた理解には有用と思われる。しかし、この種のアプローチは、完了していない進行中の議論については不向きであり、何より議論の経過の把握に対する支援とはなり難い。

また、メール分類に関する研究 (たとえば文献 [7], [8], [9], [10]) も数多く行われている。エンドユーザによる日常のメール管理からスパム抽出まで応用範囲が広く有用だが、議論の結果や経緯の把握を直接的に支援することはできない。

一方、議論の流れを抽出しようとする研究 [2], [11], [12], [13] もいくつか報告されている。返信関係などに基づいて電子メール間のつながりを再現することを目的としており、個々のメールに含まれる記述のうち議論に関わるものがどれかを抽出することは考慮していない。そのため、1つの電子メールに複数の話題を含む場合には対応できない。

メール中の議論に関わる記述をより詳細なレベルでとらえた議論構造抽出を試みている研究 [14], [15] も一部存在している。しかし、対象とする電子メールが限定的であったり、電子メール自体の記述方法 (たとえば返信メールの

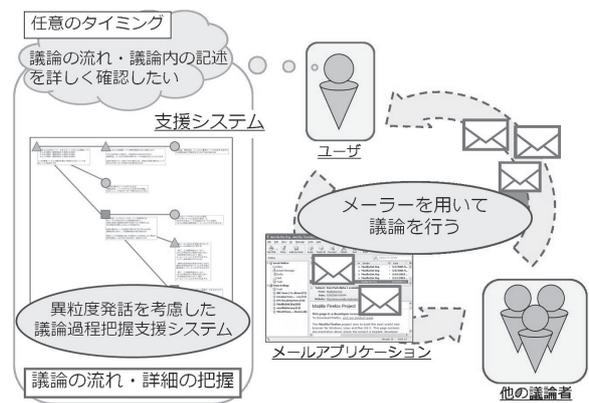


図 1 支援環境のイメージ

Fig. 1 Overview of the support environment.

記載方法) に強い制約を課していたりするため、現状では十分に有効とはいえない。

これらのことから、電子メールの記述方法に強い制約を課せずに、できるだけ通常使用の電子メールからそのまま議論過程を抽出し、議論の経過を把握するための助けとなる情報を提供しうることが必要といえる。

2.4 方針

本研究では、まず、議論ごとに電子メールをグループ化し、議論に関する記述を発話レベルで抽出する手法を開発する。次に、ある記述とそれに対する返答にあたる記述の対 (「応答記述ペア」または「応答ペア」と呼ぶ) を抽出する手法を開発する。この際、実際の電子メール議論における応答様態の分析に基づいて、典型的な記述粒度と応答レベルをパターン化する。この応答パターンに基づいて、粒度の異なる記述間の応答ペア (異粒度応答ペア) の抽出を可能とすることを目指す。さらに、異粒度応答ペア間の関係の分析に基づいて議論過程を抽出する手法を開発する。これにより、記述粒度の多様性と複雑な応答関係に対応した議論過程抽出を実現する。そのうえで、これらの手法を導入した議論過程把握支援システムを開発する。本システムは、話題ごとの電子メール管理を容易に行える機能・環境とともに、抽出した議論構造をツリー形式の視覚的情報としてユーザに提示する。これにより、ユーザが自身の保持する電子メール中の議論過程の把握を容易に行えるようにすることが、本研究の最終的な目標である (図 1)。

議論過程抽出に関する理論的側面から、応用システムの実践的側面まで及ぶ本研究の性格ゆえ、側面・要素ごとに慎重に検証を重ねることが重要と考えられる。そこで本論文の目的は、新たな議論過程抽出方法を明らかにし、実際のメールデータを用いた実験を通して、理論的手法の側面での有効性を検証し、応用システムへの導入に耐えうるかを議論することである。

3. 記述形式の多様性を考慮した議論過程抽出手法

本研究の抽出手法は、以下に示す3工程で構成されている。工程を分けてデザインすることで、実践的・継続的な手法の改良が可能なフレームワークの特性を持たせている。

Step1 話題ごとの電子メールのグルーピング

Step2 発話レベルでの議論に関する記述の抽出

Step3 異粒度応答ペアと議論構造の抽出

3.1 話題ごとの電子メールのグルーピング

電子メールを話題ごとにグルーピングする。ここでは、電子メール間の返信関係と引用関係の2つに着目することで、話題ごとのグルーピングを実現する。

まず、電子メール群から返信関係を抽出する方法として、電子メールに付与されるヘッダ情報の解析を行う。電子メールには本文以外に、1通1通に対して特殊な情報が記載された領域がある。この特殊な領域に記載された情報のことをヘッダ情報と呼ぶ。本研究では、主に電子メールの送信先を特定するために、Message-ID, In-Reply-To, References の3つの値を解析する(表2)。

最初に、軸となる電子メールを選定し、ヘッダ情報から軸の電子メールに対する返信の電子メールを特定する。特定した返信の電子メールを新たな軸として、同様の処理でさらに返信の電子メールを特定する。そして返信の電子メールがなくなるまで処理を続ける。この解析により特定した電子メール群を返信関係による同一話題のグループとする。ヘッダ情報に基づく軸メール、返信メールの具体的な解析は以下の手続きを踏む。まず、軸の電子メールは、In-Reply-To, References いずれの値も持たないものを優先して選ぶ。選定した電子メールのMessage-IDの値を取得する。次に、Message-IDとIn-Reply-Toの値から返信の電子メールを特定する。軸の電子メールのMessage-IDとIn-Reply-Toの値が一致する電子メールが返信のメールとなる。また、In-Reply-Toの値を持たない電子メールであってもReferencesの値から、返信先となる電子メールを特定できる場合もあるので、こちらの値も参照する。

次に、電子メールの本文から引用符(">"など)付きの記述を解析することで、電子メール間の引用関係を抽出する。具体的には、まず、解析対象の電子メールの本文を先

表2 返信関係の抽出に用いるヘッダ情報

Table 2 E-mail headers used for extraction of reply-relationship.

フィールド名	フィールドの値
Message-ID	電子メール1通1通に付与される固有の番号
In-Reply-To	返信先となる電子メールのMessage-ID値
References	In-Reply-To値とそれまでのReferences値の累積

頭行から1行ずつ読み込み、その先頭語句に引用符が記述されている文章を探す。ここで、引用符が記述された文章を引用文と呼ぶ。この引用文を基に、解析対象以外の電子メール群の本文を1行ずつ読み込み、引用文と記述内容が一致する文章を含むメールを特定する。引用文を含む電子メールと、引用文の記述内容と一致する記述を含む電子メールは引用関係を持っている。そして、これらの電子メールを同一の話題に関するものとしてグルーピングする。ここで、引用文と同一の記述を他のメールから抽出する際に、議論参加者ごとのスタイルや使用するメールソフトの設定によっては、改行位置が変化することがある。また、引用を途中省略するケースもある。本手法では、これらが、1つの引用発話の後半に行くにつれてより発生しやすいことに注目し、引用発話の1行目に着目して、それと一致する発話を走査することにより、引用文と同一の記述を抽出する。これにより、改行位置の変更や引用の中略などに対応している。ただし、引用発話の1行目ですら改行位置が変わっている場合には、現状では対応していない。この一致判定についてはチューニングの余地がある。

本論文では、馴染み深い引用符の1つとして、">"を例にあげた。しかし、本手法では、引用符と見なす記号を事前に指定する形をとっており、引用符を">"に限定するわけではない。したがって、たとえば全角・半角両方の">"を引用符として扱うことも可能である。

本手法は、2種類のグルーピング手法を組み合わせることで、抽出洩れの少ない電子メールグルーピングを実現する。

3.2 発話レベルでの議論に関する記述の抽出

話題ごとにグルーピングした電子メール群から、発話レベルで議論に関する記述を抽出する。本研究では、電子メール本文を構成する1文以上の意味をなす文章の塊を発話と定義する。メールレベルよりも詳細な粒度で的確に議論に関する記述を把握するために、電子メール本文から発話を抽出する。

まず、文どうしの切り分けを行う。1行以上の空行による切り分けを基本とする。さらに、電子メール本文中に記述される、句点(“。”や“.”)、疑問符(“?”)、読点(“、”や“、”)などの文の完結、前後の記述の接続を意味する特徴的な記号、引用符(">")や接続詞などの手がかり語を基に、それぞれの発話の範囲を特定したうえで記述を抽出する(図2)。

次に、抽出した発話から、議論に関するもののみを抽出する。発話には、議論に関するもの以外にも、様々なものが存在している。ここで、議論に関する発話を有効発話、それ以外を無効発話と呼ぶ。本研究では、実際の電子メール本文の観察に基づき、以下の3種類の代表的な無効発話を対象に除去を行う。

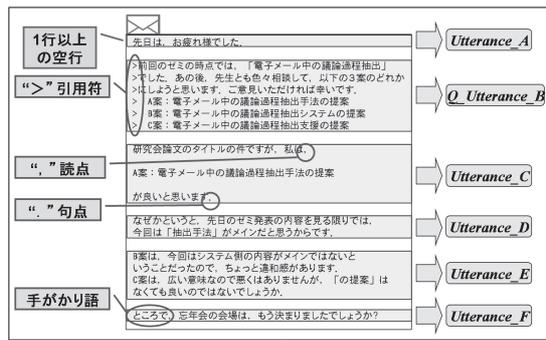


図 2 発話レベルでの記述の抽出

Fig. 2 Extraction of the descriptions in utterance level.

- メール冒頭に記述される挨拶文・名乗り表現
- 本文最下行の全文引用文
- 所属を示すシグネチャ文

挨拶文・名乗り表現は、電子メール本文中の先頭行に記述される「こんにちは」、「お世話になっております」などの挨拶文と、「〇〇(人名)です」などの自身の氏名を名乗る表現を指す。一般に定型文として記述されるこれらの表現を、電子メール本文の1~5行目を対象に固有フレーズ集によるマッチングを行うことで除去する。

全文引用文は、返信の電子メールを作成する際に、メール本文の最下行に挿入された「返信先の電子メール」を指す。この全文引用文は、メーラの設定によって自動で挿入される場合が多いため、返信内容との直接的な関係のない無効発話を多く含む。そこで、電子メール本文の最後尾を対象に、引用符(">"など)のマッチングにより除去する。

シグネチャは、送信者の名前、住所、所属機関などを表す署名情報である。このシグネチャは、一般に議論内容との関連性が低い。そこで、全文引用文を除いた電子メール本文の最後尾を対象に、以下の3つのポイントに基づいて判定・除去する。

- 記号のみで構成される行の有無
- 各行末尾の句読点の有無
- 判定対象の行数

3.3 異粒度応答ペアと議論構造の抽出

電子メール議論は、複数の議論メンバによる返信行為の積み重ねによって構成される。この電子メールのやりとりの中に含まれるある記述とそれに対する返信にあたる記述の対(応答ペア)を、返信メール中での引用に着目した解析によって詳細なレベルで抽出する(図3)。

【処理1】まず、無効発話を除去した発話群を「引用」を手がかりに分析し、返信発話をその返信先とともに洗い出す。一般に、引用文に続けて、それに対する返信を記述することが多い。しかし、実際の議論における発話間の関係は複雑である。つまり、1発話に対して1発話が返答するとは限らない。たとえば、1発話に対して、3発話をもつ

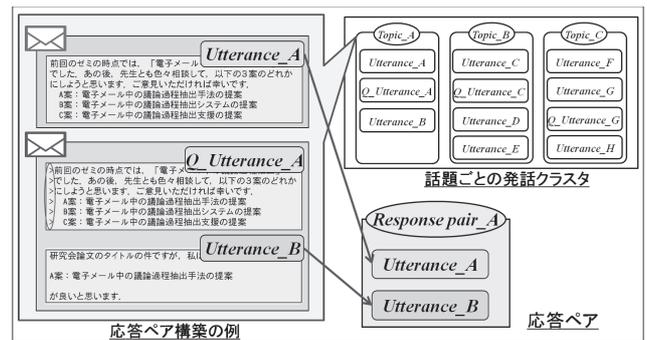


図 3 応答記述ペアの抽出

Fig. 3 Extraction of response pair.

て返答する場合がある。また、1通の電子メール全体に対して、1発話で返答する場合もある。この場合は、返答発話が引用文より前に位置する。このように、多様な応答記述形式が混在する実際の電子メールデータから、応答ペアをうまく取り出すことは、既存手法では困難であった。本研究では、返答時の記述粒度の多様性を許容しうる応答記述ペア抽出手法を開発する。

筆者らは、実際の電子メール議論を丁寧に観察し、応答記述形式を6つの典型パターン(応答スタイル)に整理した(表3)。ここで、引用文とそれに対する応答発話は、応答スタイルによって特有の出現様態をとる。

【処理2】この点に着目し、発話の出現様態の解析によって、個々の応答記述について、応答スタイルを判別する。まず、電子メール本文の有効発話を上から順に引用符に着目して検査し、引用には1、返答には0を与える。これにより、引用文とそれに対する返答発話の出現様態を表すビット標記(発話セット)を得る。そのうえで、個々の応答発話の出現様態が、どの応答スタイルに該当するかを判定することで、応答・被応答記述の粒度を把握する。たとえば、対象電子メール中の3つの発話が返信先の電子メール中の異なる3つの発話に対してそれぞれ応答している場合や、対象電子メール中のすべての発話が返信先の電子メール全体に対する応答である場合、などを区別できる。

【処理3】この応答スタイルの判別に基づいて、個々の応答記述から小さなツリー(応答ペア)を生成する(図4)。具体的には、応答・被応答記述をノード、記述間の返答関係をエッジで表す。このような応答ペアを表現するために、3種類のノード(単独発話、複数発話、電子メール)を導入する。これにより、既存手法では不可能であった応答記述形式の混在への対応が可能となる。

【処理4】次に、任意の応答ペアを2つ取り上げ、同じ記述を含むか否かを診断する。同じ記述を共有している場合には、それら2つの応答ペアを共通記述を接点としてつなぎ合わせる。この際、同じ記述を含むが、粒度が異なるノードが存在するケースもある。この場合は、共通記述を有する2つの比較対象ノードのうち、粒度が小さい方のノード

表 3 応答記述形式の典型パターン
Table 3 Typical patterns of response style.

応答形式	被返答側の記述粒度	返答側の記述粒度	発話セット
1 発話対 1 発話	1 発話	1 発話	[1, 0]
n 個の発話対 1 発話	複数発話	1 発話	[1, ..., 1 _n , 0]
1 発話対 n 個の発話	1 発話	複数発話	[1, 0, ..., 0 _n]
n 個の発話対 m 個の発話	複数発話	複数発話	[1, ..., 1 _n , 0, ..., 0 _m]
電子メール対 1 発話	電子メール	1 発話	[0]
電子メール対 n 個の発話	電子メール	複数発話	[0, ..., 0 _n]

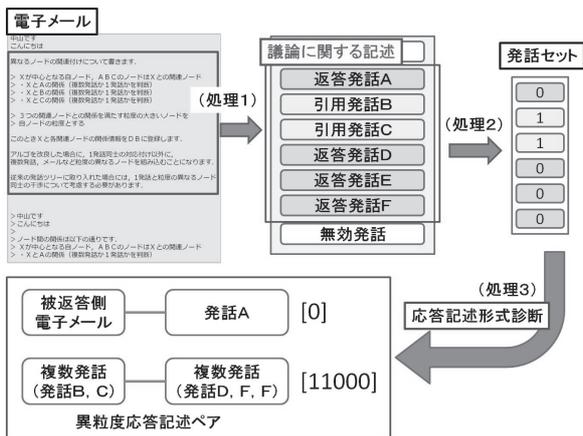


図 4 異粒度応答ペアの抽出

Fig. 4 Extraction of response pair in various granularities.

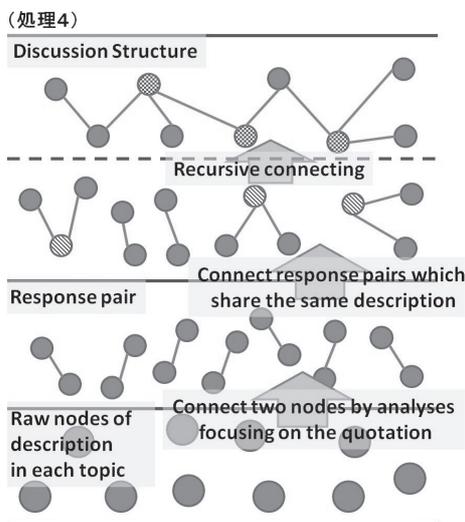


図 5 議論構造の抽出

Fig. 5 Extraction of discussion structure.

の記述すべてがもう一方のノード中に含まれるか否かで、応答ペアどうしをつなぎ合わせる接点を判定する。一方、比較対象の2つのノードの粒度が同じ場合（つまり、1 発話対 1 発話、 n 個の発話対 n 個の発話、電子メール対電子メール）には、記述内容が互いに一致するか否かを検査する。応答ペアどうしの共通記述のつなぎ合わせを再帰的に行うことで、大きいツリー（議論構造）を生成する（図 5）。なお、これらの抽出手法は、3 名以上の議論にも対応して

いる。その場合、議論参加者の増減にともない、議論に関わるメールの宛先が変化する場合が考えられる。解析の最初の段階で議論参加者を抽出し、それを基にして議論に関わるメール・記述を追跡する形の方法では、参加者の増減に対応できない場合がある。これに対して、提案手法では、当初抽出した議論参加者に頼って当該議論に関わるメールを抽出するのではなく、メールのヘッダ情報 (In-Reply-To, References)、および、メール本文中での引用関係という、いずれも議論参加者（メールの宛先）に依存しない情報の解析による「議論に関わるメールのグルーピング」を議論過程抽出の最初の段階に位置づけている。したがって、議論参加者の増減（メール宛先の変化）にも対応している。

また本手法は、メールのヘッダ情報や、引用文と返答文の関係など、構造的特徴の解析を主としている。そのため、使用言語に依存せず、基本的には日本語以外にも対応している。ただし、日本語以外の新たな言語に適用する場合には、Step2 発話レベルでの議論に関する記述の抽出における特徴フレーズの指定など、ある程度の準備は必要である。

4. 議論過程把握支援システム

本研究で開発しているシステムは、ユーザが蓄積している電子メールを本手法に基づいて解析することで、リアルタイムで行われている議論や過去に行われた議論から、その具体的な流れを抽出し、視覚的に提示するものである。主に、抽出アルゴリズムに基づくメールデータの解析処理は Java で、GUI などの視覚的提示に関する部分は Flash で実装されている。メールの作成・送受信は、ユーザがふだん使用しているメーラなどをそのまま使用し、議論過程を把握しようとする場面で、本支援システムを活用するスタイルを想定している。

次に、本支援システムの利用方法を示す。システム上には、メールを話題ごとに管理するリストと、ユーザが把握したい話題を視覚化するスペースが設置されている（図 6）。ユーザは、話題をまとめたリストの中から、把握したい話題を選択することで、視覚化スペースに議論の流れが視覚的に提示される。視覚化スペースに表示されたノードが議論に関する記述を表し、記述内容をまとめたテキストが付与されている。エッジによるノード間のつながりが、議論に関する記述どうしの直接的な話のつながりを示す。ユー

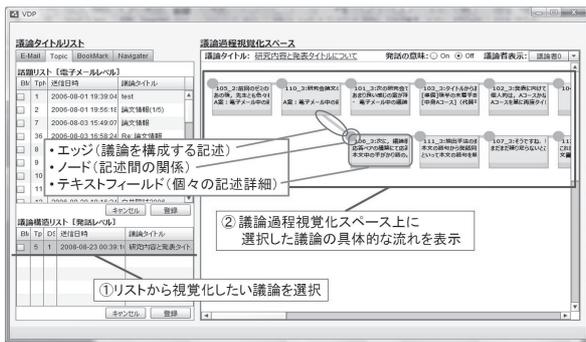


図 6 支援システムのインターフェース
Fig. 6 Interface of the support system.

ずは視覚化された議論過程から、表示されたノードの記述内容やノード間のつながりを見ることで、議論内容の把握ができる。

なお、それぞれのフェーズにおいてユーザは、メールの分類など支援システムによる処理結果を、簡単なマウス操作で適宜補正することができるデザインをとる。ゆえに、各段階でシステムが提示する情報は、これらユーザによる情報管理・補正を容易に行えるようにするための下地にあたる。

5. 実験

5.1 実験概要

本研究で開発した議論過程抽出手法の有効性の検証、課題抽出を目的として実験を行った。まず、実験協力者2名が、それぞれ自身のメールアプリケーションに蓄積されている電子メール1,000件を対象として、以下1)~3)の作業を行った。

- 1) 議論を構成するメール集合を抽出する。
- 2) 抽出した各メール集合について、議論に関係する記述をピックアップする。このとき、記述の実態に応じて、電子メール、連続する複数発話、単一の発話のいずれかでピックアップする。
- 3) 並行して、2)で抽出した記述間の関係を線で結ぶ。

なお、1)の作業には、通常使用しているメールアプリケーションの検索・並べ替え機能などを使用してよいものとし、2), 3)は、MS-Officeファイル上に当該記述を張り付け、直線描画機能で線を結ぶ形で実施した。つまり実験協力者は最終的に、議論ごとにそれを構成する記述とその間の関係を示すツリー(議論ツリー)を作成した。この際、メール閲覧の順番など、議論を構成する記述の検索に関わる作業手順は、実験協力者のやりやすい方法でよいこととした。

一方、提案手法を実装したプログラムを用いて、同じメールデータを対象に1)~3)を実施した。その後、実験協力者による結果に対する、提案手法を実装したプログラムによる抽出の一致状況を、発話などを1つ1つ実際に確認して、精査した。

具体的には、1)「議論を構成するメール集合の抽出」については、以下の方法で適合率、再現率をそれぞれ算出した。 $MAIL_p$ は、提案手法を実装したプログラムが抽出した議論に関するメールの集合、 $MAIL_u$ は、協力者が抽出した議論に関するメールの集合をそれぞれ表す。 n は件数(この場合、メール数)を表す。ただし、今回の実験では、議論を区別して評価を行った。したがって、プログラムが実験協力者と同じメールを抽出しても、どの議論を構成するメールかの推定が誤っていれば不正解とした。

$$\text{適合率} = \frac{n(MAIL_p \cap MAIL_u)}{n(MAIL_p)}$$

$$\text{再現率} = \frac{n(MAIL_p \cap MAIL_u)}{n(MAIL_u)}$$

2)「議論に関わる記述のピックアップ」については、実験協力者・提案手法(プログラム)ともに「発話単位で抽出する」ことを前提としている。1つのメール中のすべての発話をピックアップする場合は、1件のメールごと抽出することに相当する。評価は、提案手法を実装したプログラムが抽出した発話のうち、協力者が抽出した発話と一致するものの確認を、抽出対象が複数発話の場合も含め、あえて発話1つ1つの単位で行った。実際の議論過程把握に照らして、複数発話(の中の発話)をまったく抽出できなかった場合と、複数発話のうちの一部が欠損しているような場合では価値が異なることを考慮し、この方法とした。そのうえで、以下の方法で適合率、再現率をそれぞれ算出した。 $UTTE_p$ は、提案手法を実装したプログラムが抽出した議論に関する発話の集合、 $UTTE_u$ は、協力者が抽出した議論に関する発話の集合をそれぞれ表す。 n は件数(この場合、発話数)を表す。

$$\text{適合率} = \frac{n(UTTE_p \cap UTTE_u)}{n(UTTE_p)}$$

$$\text{再現率} = \frac{n(UTTE_p \cap UTTE_u)}{n(UTTE_u)}$$

抽出した発話の一致確認においては、「明らかに同じ発話を抽出しているが、抽出方法にわずかな差がある場合(たとえば、同じ発話を抽出しているが、発話の末尾の文字列をわずかに抽出していないなど)」には、一致と判断した。発話単位での抽出を前提としているため、上記以外で、抽出粒度がプログラムと協力者の間で異なる状況は生じなかった。なお、2)についても、1)と同様に、議論を区別して評価を行った。

3)「議論を構成する記述間の関係抽出」については、関係(エッジ)の両端の記述の粒度(1発話と1発話、1発話と複数発話など)にかかわらず、すべて1つの関係として評価した。この際、エッジの両端については、プログラムが抽出したエッジ端点と協力者によるものが同一の場合、および、プログラムが抽出したエッジ端点が協力者によるものを完全に含む場合(たとえば、協力者によるもの:[発話

a]-[発話 bc], プログラムによるもの:[発話 a]-[発話 bcd]) は, 当該エッジ (関係) を一致と判断した. 逆に, プログラムが抽出したエッジ端点が協力者によるものの一部のみになっている場合 (たとえば, 協力者によるもの:[発話 a]-[発話 bc], プログラムによるもの:[発話 a]-[発話 b]) には, 当該エッジを不一致と判断した. これは, 同じ「ある関係 (完全一致ではないもの) おおむね抽出できている場合」であっても, 実際の議論過程把握に照らすと, 端点にあたる発話を余計に (広く) 抽出している場合は比較的影響が少ないのに対し, 端点にあたる発話の一部欠損している場合は影響が大きいと考えられるためである. そのうえで, 以下の方法で適合率, 再現率をそれぞれ算出した. $EDGE_p$ は, 提案手法を実装したプログラムが抽出した関係 (エッジ) の集合, $EDGE_u$ は, 協力者が抽出した関係の集合を, それぞれ表す. n は件数 (この場合, 関係数) を表す.

$$\text{適合率} = \frac{n(EDGE_p \cap EDGE_u)}{n(EDGE_p)}$$

$$\text{再現率} = \frac{n(EDGE_p \cap EDGE_u)}{n(EDGE_u)}$$

なお, これまでこの種の抽出手法については, 実際の電子メールデータを用いた踏み込んだ精度検証は, その作業コストゆえにあまり行われてこなかった. 本研究では, 1,000 件という比較的大きな規模の実際のメールを用いて, 相当の時間をかけて丁寧な検証を行った.

ここで, 本手法には, 議論に関するメールや発話をなるべく広く抽出することが求められる. これは, 議論に関係する可能性のある発話群から, ユーザが余計なものを除く形の補正が現実的と思われるためである. もし逆に, 議論に関係する可能性のある発話群として提示された情報が実際に必要なものばかりであったとしても, 候補として抽出されないものが多すぎる場合, ユーザがそれを自力で大量の生メール群から探し出すことは困難だからである. 本手法には, いわば「石を拾うことはあっても玉を捨てない」性格, つまり, 再現性が重要と考えられる.

5.2 結果と考察

表 4 に, 1)「議論を構成するメール集合の抽出」の結果を示す. ここでは, 実験協力者が抽出したメール集合に対する, 提案手法を実装したプログラムによるメール集合抽出の精度を算出した. 適合率・再現率ともに高い値が得られている. さらに各メール集合について実際に観察したと

表 4 議論を構成するメール集合の抽出結果

Table 4 Results of extraction of e-mails which build a discussion.

実験協力者	適合率	再現率
協力者 A	0.93	0.92
協力者 B	0.89	0.89

ころ, 実験協力者によるメール集合抽出に照らして, 提案手法は返信関係の解析で抽出しきれないメールを, 引用関係の解析によりうまくとらえることができていた. これらのことから, 本手法が話題ごとのグルーピングのアプローチとして有効であることが分かった.

表 5 に, 2)「議論ごとのメール集合からの議論に関わる記述のピックアップ」の結果を示す. ここでは, 実験協力者が抽出した議論に関わる記述 (実験で作成した議論ツリーのノード) 群に対する, 提案手法を実装したプログラムによる記述群抽出の精度を, 発話単位で算出した. まず, 再現率については良好な結果が得られている. 一方, 適合率は他の部分と比較すると低めの値となった. 実際の状況を確認したところ, 提案手法は候補発話を広くとる方針なのに対し, 人 (実験協力者) はメール中からピンポイントで議論に関わる記述を選んでとる傾向があることが影響していることが分かった. ただし, 今回の実験では, 「全発話から, 議論に関わる発話を抽出する (議論に関わる発話を抽出できれば正解)」だけではなく, 「全発話から, 議論に関わる発話を (どの議論に関わるのかを区別して) 抽出」している. そのため, 議論に関わる発話を抽出できたとしても, 該当する議論を間違えている場合には不正解と判定している. 今回の実験では, このように踏み込んだ評価方法をとっているため, 一般にトレードオフの関係にある適合性・再現性について, どちらを重視するか, より明確な方針が求められる. ここでは, 5.1 節で述べたとおり, 再現性を重視すべきである. そのことに照らして, 初期段階としてはおおむね良好な結果と考えられる.

表 6 に, 3)「議論を構成する記述間の関係抽出」の結果を示す. ここでは, 実験協力者が抽出した記述間の関係 (実験で作成した議論ツリーのエッジ) 群に対する, 提案手法を実装したプログラムによる記述間関係の抽出精度を算出した. 適合率・再現率ともに良好な値が得られている. ここで, 想定する応用システムを用いたメール管理では, 各段階で適宜補正が行われるのに対し, 今回の実験では 1) におけるプログラムの抽出結果をそのまま 2) に使用して

表 5 議論に関わる記述のピックアップの結果

Table 5 Results of picking up of descriptions which build a discussion.

実験協力者	適合率	再現率
協力者 A	0.28	0.88
協力者 B	0.25	0.86

表 6 議論を構成する記述間の関係抽出の結果

Table 6 Results of extraction of relationships among descriptions which build a discussion.

実験協力者	適合率	再現率
協力者 A	0.70	0.63
協力者 B	0.70	0.74

表 7 実験に用いた電子メールデータの概要

Table 7 Outline of e-mail data used in the experiments.

実験協力者	メール総数	発話総数	議論数	議論に関する 発話の数
協力者 A	1,000	21,112	13	161
協力者 B	1,000	10,918	44	391

いることに注意されたい。応用システムを活用する段階では、2), 3)の結果がおのずと良い方向へ変化することになる。これらを考え合わせると、応答の診断に基づく応答ペア、議論ツリーの抽出手法についても、現段階では良好な結果と考えられる。

なお、3)「議論を構成する記述間の関係抽出」の抽出精度は、2)「議論に関わる記述のピックアップ」の抽出精度の影響を受ける。ただし、今回の実験では、5.1節で述べたように、2), 3)は、実際の議論過程把握に照らした趣旨をふまえ、それぞれの方法で評価している。そのため、ここでは、3)「議論を構成する記述間の関係抽出」の適合率(表6)は、2)「議論に関わる記述のピックアップ」の適合率(表5)を必ずしも下回るとは限らない。

実験に用いた電子メールデータの概要を表7に示す。今回の実験で抽出された応答ペアの分布は、「電子メール対 n 個の発話」が約90%、「1発話対1発話」「1発話対 n 個の発話」「電子メール対1発話」が、それぞれ3%前後であった。記述形式が比較的シンプルであろう実験協力者(学生)であっても、多様な形式の応答ペアが存在し、それらを実際に抽出できた意義は大きいものと考ええる。一方、「 n 個の発話対1発話」「 n 個の発話対 m 個の発話」は抽出されなかった。より多様な記述形式を用いる可能性のあるユーザを含め、さらに検証を重ねていくことが重要と考える。

また、今回の実験では、提案手法を実装したプログラムでは、全角“>”と半角“>”の両方を引用符に指定した。一方、対象とする電子メールについては、引用符や引用形式などについて特定のものを選ぶことはせず、実験協力者から提供された電子メールをそのまま用いた。結果的に、“>”以外の引用符を使用している箇所、表3に示したものの以外の応答形式は見受けられなかった。これは、今回の実験協力者が学生だったこともあり、ごくポピュラーな引用符の使用・引用形式であったものと考えられる。より複雑な議論を行うことが多い大学教員のような議論者の場合を含め、引用符指定などの準備レベルと提案手法による抽出結果の兼ね合いなどについても検証を重ねることが重要と考える。

なお、今回の評価実験では、対象とする議論の種類について特に指定はしなかった。実験協力者がともに大学院生であったこともあり、結果的に、実験協力者自身の研究に関する共同研究者とのやりとり、研究室での行事に関する他の学生との連絡・意見調整など比較的客観的な議論が

多くを占めていた。今後は、感情的なやりとりを含め、性質の異なる様々な議論について入念に検証を重ねていくことが重要だと考える。

今回の実験は被験者数などが限定的なものであり、これらの結果だけからただちに提案手法の有効性を確認したとまではいえないものの、全体として十分良好な感触といえる。また、実験協力者による作業には、2名ともに10時間程度を要した。このことから、本研究が目指す踏み込んだ議論過程把握は、支援なくては大変困難なものであることが分かり、提案手法の意義を垣間見ることができる。

6. むすび

本論文では、電子メール中の議論過程抽出手法の開発について述べた。実際の電子メール議論における応答様態の分析により典型的な記述粒度と応答記述形式をパターン化し、それに基づいた異粒度応答ペアと議論構造の抽出方法を示した。また、実際の電子メールデータを用いた実験を行い、提案手法の有効性について良好な感触を得た。同時に、人の発話選別における傾向などいくつかの貴重な知見を得ることができた。

今後は、議論が行われる業務などの様態、議論者の立場や専門分野など、できるかぎり多様な対象について提案手法を適用し、検証を重ねることが重要と考える。また、提案手法の応用システムへの導入とその試用を進め、想定外のケースの確認、利用者の主観面などを含めた実践の評価を行い、議論把握支援としての有効性検証につなげたい。

謝辞 本研究の一部は、科学研究費補助金若手研究(A)(No.25702007)の支援を受けている。検証作業などに親身なるご協力をいただいた福島大学・山崎辰也氏(現在、株式会社沖データ)、安齋洋行氏(現在、アルパイン情報システム株式会社)、森田裕彬氏、実験などにご協力いただいた中村研究室の皆さんに深謝いたします。

参考文献

- [1] 伊地知宏, 倉部 淳: メールでの話題の把握を支援するソフトウェア, IPA 未踏ソフトウェア創造事業成果報告書(2003), 入手先(<http://www.ipa.go.jp/files/000005584.pdf>) (参照 2014-04-10).
- [2] 松村真宏, 加藤 優, 大澤幸生, 石塚 満: 議論構造の可視化による論点の発見と理解, 日本知能情報ファジィ学会誌, Vol.15, No.5, pp.554–564 (2003).
- [3] Nakayama, H., Onuma, R., Kaminaga, H., Miyadera, Y., Yokoyama, S. and Nakamura, S.: Extraction Methods of E-mail Discussion Processes Considering Diversity of Description Granularity and their Complicated Relationships, Proc. 8th International Conference for Internet Technology and Secured Transactions, pp.563–568, IEEE (2013).
- [4] 稲葉晶子, 岡本敏雄: 協調活動における議論支援戦略, 電子情報通信学会技術研究報告, Vol.98, No.35, pp.77–84 (1998).
- [5] 青木志乃, 長瀧寛之, 大下福仁, 角川裕次, 増澤利光: 複

数グループのオンライン議論を同時にサポートする自動助言システムの構築, 情報処理学会研究報告, Vol.2009, No.3, pp.79-84 (2009).

- [6] Hatori, J., Murakami, A. and Tsuji, J.: Multi-Topical Discussion Summarization using Structured Lexical Chains and Cue Words, *Proc. 12th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2011)*, pp.313-327, Springer (2011).
- [7] 大島裕明, 小山 聡, 田中克己: 個人文書から抽出した語彙の意味関係に基づく Web 情報検索, 情報処理学会研究報告, Vol.2005, No.67, pp.105-111 (2005).
- [8] 河重貴洋, 大島裕明, 小山 聡, 田島敬史, 田中克己: コンテキストを用いたメールの情報補完, 情報処理学会研究報告, Vol.2006, No.78, pp.329-336 (2006).
- [9] Joty, S., Murray, G. and Ng, R.T.: Supervised topic segmentation of email conversations, *Proc. ICWSM 2011*, available from <http://www.cs.ubc.ca/~rjoty/paper/ICWSM-CR.pdf> (accessed 2014-04-10).
- [10] Joty, S.R., Carenini, G. and Ng, R.T.: Topic Segmentation and Labeling in Asynchronous Conversations, *J. Artif. Intell. Res. (JAIR)*, Vol.47, pp.521-573 (online), DOI: 10.1613/jair.3940 (2013).
- [11] 松村真宏, 大澤幸生, 石塚 満: テキストによるコミュニケーションにおける影響の普及モデル, 人工知能学会論文誌, Vol.17, pp.259-267 (2002).
- [12] Dehghani, M., Shakeri, A., Asadpour, M. and Koushkestani, A.: A Learning Approach for Email Conversation Thread Reconstruction, *Journal of Information Science*, Vol.39, No.6, pp.846-863 (2013).
- [13] Dehghani, M., Asadpour, M. and Shakeri, A.: An Evolutionary-Based Method for Reconstructing Conversation Threads in Email Corpora, *Proc. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp.1132-1137, IEEE (2012).
- [14] 村越広享, 島津 明, 落水浩一郎: 電子メールを利用したコミュニケーションにおける討議ストリームの自動抽出法, 電子情報通信学会技術研究報告, Vol.98, No.558, pp.33-40 (1999).
- [15] 村越広享, 山見太郎, 島津 明, 落水浩一郎: 電子メールを利用した学習者間のコミュニケーション支援技術の開発, 教育システム情報学会誌, Vol.18, No.3, pp.308-318 (2001).



中山 祐貴 (学生会員)

2009年福島大学理工学群共生システム理工学類卒業。2011年同大学大学院共生システム理工学研究科修士課程修了。現在、同研究科博士後期課程在学中。データ工学, 情報視覚化, 議論支援, 教育工学, 等の研究に従事。

2009年本会東北支部学生奨励賞受賞。IEEE Computer Society/Education Society, 電子情報通信学会各会員。



大沼 亮 (学生会員)

2010年福島大学理工学群共生システム理工学類卒業。2012年同大学大学院共生システム理工学研究科修士課程修了。現在、同研究科博士後期課程在学中。知的活動過程における文脈情報抽出・組織化, データ工学, 教育工学, 等の研究に従事。

IEEE Computer Society/Education Society, 電子情報通信学会各会員。



神長 裕明 (正会員)

1979年東北大学工学部通信工学科卒業。1989年同大学大学院工学研究科博士課程修了。同年東北大学大型計算機センター助手。1991年山形大学工学部助手, 同年10月同大学工学部講師, 1995年福島大学教育学部助教授を

経て, 2004年より福島大学理工学群共生システム理工学類教授, 現在に至る。工学博士。ソフトウェア工学, プログラミング・ソフトウェア設計を対象とした演習支援システム, 教育工学, 等の研究に従事。IEEE Computer Society, ACM, 電子情報通信学会, 日本知能情報ファジィ学会, 日本ソフトウェア科学会, 人工知能学会, 日本産業技術教育学会各会員。



横山 節雄

1968年早稲田大学理工学部応用物理学卒業。1974年同大学大学院博士課程修了。東京学芸大学助手, 助教授, 教授を歴任。現在, 同大学名誉教授。理学博士。1992~1993年までスタンフォード大学客員研究員。教育工学,

情報技術の教育領域への適用, 等の研究に従事。電子情報通信学会教育工学研究会専門委員長, 日本教育工学会理事等を歴任。



宮寺 庸造 (正会員)

1984年東京電機大学理工学部数理学科卒業。1986年同大学大学院理工学研究科数理学専攻修士課程修了。同年同大学理工学部情報科学科助手。1999年東京学芸大学教育学部数学・情報科学科講師。2001年同大学助教授, 2009年同大学教授, 現在に至る。博士(理学)。2001~2002年まで文部科学省メディア教育開発センター客員助教授併任。2012~2014年まで湖南師範大学客員教授。プログラム言語処理系, アルゴリズム, 情報視覚化, 教育工学, 等の研究に従事。IEEE Computer Society, ACM, 電子情報通信学会, 日本ソフトウェア科学会, 人工知能学会, 教育システム情報学会, 日本教育工学会各会員。



中村 勝一 (正会員)

1998年東京学芸大学教育学部情報環境科学課程教育情報科学専攻卒業。2003年会津大学大学院コンピュータ理工学研究科博士課程修了。同年東北学院大学教養学部情報科学専攻講師。2004年福島大学理工学群共生システム理工類(数理・情報学系)講師, 2008年同大学准教授, 現在に至る。博士(コンピュータ理工学)。データ工学, 情報視覚化, Web Intelligence, 教育工学, 等の研究に従事。IEEE Computer Society, ACM, 電子情報通信学会, 教育システム情報学会各会員。