特集論文 「近未来チャレンジ 2010」

Total Environment for Text Data Mining

テキストデータマイニングのための統合環境

砂山 渡 Wataru Sunayama
広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University
sunayama@hiroshima-cu.ac.jp

高間 康史 Yasufumi Takama
首都大学東京システムデザイン学部
Faculty of System Design, Tokyo Metropolitan University
ytakama@sd.tmu.ac.jp

Danushka Bollegala ダヌシカ ボレガラ
東京大学大学院情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
danushka@iba.t.u-tokyo.ac.jp

西原 陽子 Yoko Nishihara
東京大学大学院工学系研究科
Graduate School of Engineering, The University of Tokyo
nishihara@sys.t.u-tokyo.ac.jp

徳永 秀和 Hidekazu Tokunaga
香川高等専門学校
Kagawa National College of Technology
tokunaga@t.kagawa-nct.ac.jp

串間 宗夫 Muneo Kushima
宮崎大学医学部附属病院医療情報部
Medical Informatics, University of Miyazaki Hospital
kushima@fc.miyazaki-u.ac.jp

松下 光範 Mitsunori Matsushita
関西大学総合情報学部
Faculty of Informatics, Kansai University
mat@res.kutc.kansai-u.ac.jp

keywords: text data mining, total environment, data visualization, graphical user interface

Summary

In this challenge, we develop and distribute an integrated environment to flexibly combine multiple text mining techniques. Text mining techniques include numerous tasks such as salient sentence extraction, keyword extraction, topic extraction, textual coherence evaluation, multi-document summarization, and text clustering. Although tools that individually perform one or more of the above-mentioned tasks exist, it is difficult to integrate and activate multiple tools for a particular task. We attempt to provide the flexibility to integrate numerous tools that exist in the community in our proposed text mining environment. Users can use a customized version of the proposed text mining environment for their specific tasks, thereby concentrating solely on their creative work.

1. はじめに

本チャレンジ (TETDM，テトディーエム) では，複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境を構築し，電子テキストを扱う多くのユーザの，創造的活動を支援するツールの提供を目指す．

テキストマイニングと呼ばれる研究には，「重要文抽出」「キーワード抽出」「トピック抽出」「テキストの一貫性評価」「複数文書要約」「テキストクラスタリング」などさまざまな課題があり，すでに多くの研究成果も世の中で発表されてきている．しかし，それぞれの技術を利用するためのシステムやツールは，各研究者が独自に構築することが多く，また論文用の試験的なシステムとなっていたりするため，実際に世の中で使われる技術はごく一部に限られてしまっている．

また，情報を多角的に分析したいユーザは，複数のテキストマイニング技術を用いたいと考える．各研究者が配布用のシステムを提供していた場合でも，複数の技術を併用するためには，それらのシステムを各方面から別個に入手した上で，システム間のデータの受け渡しや結果の比較のために，手作業でフォーマットを整えたり，新たなインタフェースを独力で構築する必要が生じる．これらのことは，単に手間がかかるというだけでなく，直感的に試行錯誤を繰り返しながら知見を得る創造活動の妨げになる．

そこで，既存また将来の研究成果によるテキストマイニング技術を，１つのシステム内のモジュールとして扱うことができ，ユーザの選択したすべてのモジュールを連動して動作させられる環境を構築し，それを無償ツールとして公開することを目指す．これにより，先の問題点を解決する以下の効果が見込まれる．

- 複数の技術を用いたいユーザの環境が整えられ，ニーズに応じたモジュールを選択した上で，分析作業に集中することができる．
- 試験的なものを含む多くのシステムやツールが集められるため，多くの技術の実用化や再利用が見込まれる．

また，テキストマイニング技術を開発する研究者の利点として，新しい技術の開発を促進できる次の効果が見込まれる．

- 関連技術を容易に収集することができ，開発技術と関連技術との比較検討や機能拡張が容易になる．
- 各研究者が研究成果を一つのモジュールとして配付することを意識できるため，研究の高いモチベーションの維持につながる．

本チャレンジでは最終的に，複数のテキストマイニング技術を併用して分析を行う中で，当たり前の結果だけではなく，頻度が低くても価値の高いデータやパターンにも気づくことができ，複数の要因が複雑に絡み合ったデータの背後に隠れた因果関係を推測して知識創発につなげられる環境を目指す．

以下本論文では，2 章で TETDM チャレンジの課題と計画について述べ，3 章で構築する統合環境の構成について述べる．4 章で TETDM チャレンジの位置づけを明確にし，5 章で本チャレンジの社会や研究分野への貢献事例について述べ，最後に 6 章で本論文を締めくくる．

2. TETDM チャレンジの課題と計画

TETDM では，複数のテキストマイニング技術を柔軟に組み合わせて使える統合環境を構築し，作成した環境，および環境内で選択的に使用できるモジュールをダウンロードできる Web サイト (図 1) を立ち上げることを目指す．以下で，TETDM でチャレンジする課題と，達成に向けての計画について述べる．

2·1 TETDM チャレンジ

本節では，TETDM で達成したい 3 つのチャレンジとその実現に向けての課題について述べる．

§1 チャレンジ１：幅広い利用者と開発者の参入

ユーザと開発者の，利用と開発のしきいを可能な限り下げて，幅広い利用を見込める環境を構築する．具体的な数値目標として，100 万人のユーザと，1000 以上のモジュールが集められる環境を想定する．ユーザ数 100 万という数値の根拠は，各種商品が 100 万個売れればミリオンセラーと呼ばれ，ヒット商品として扱われることと，現在の SNS 上のアプリケーションにおいても，人気上位のものは 100 万を超えるユーザ数があることによっている．またモジュール数 1000 以上としたのは，広くテキストデータマイニングに関わる国内の研究者，開発者が，平均して一人一つのモジュールを作ることを想定し，幅広く簡易に開発が可能な場合に達成し得る値として算定している．

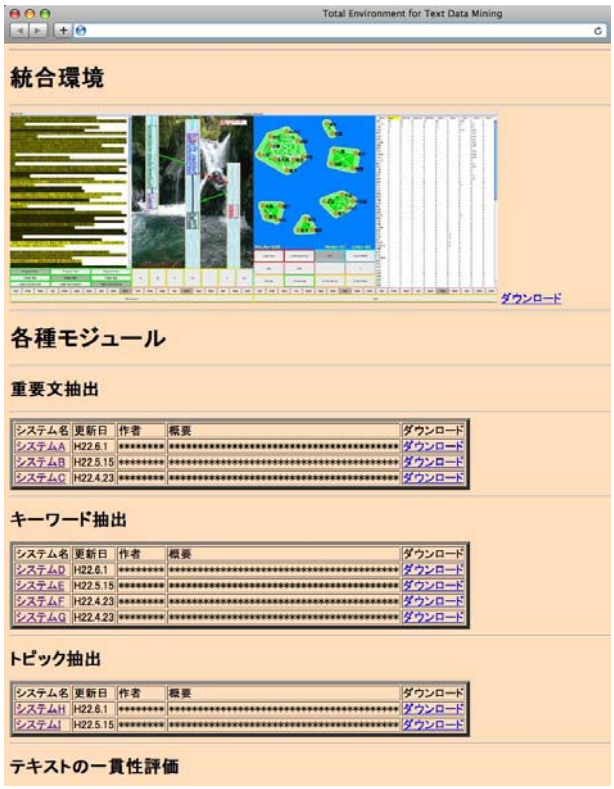

図 1 統合環境とモジュールのダウンロードサイト（イメージ）

これを達成するための必要条件として，ユーザ側の条件としては，ユーザが平易に使用できることと，また卑近なニーズに答えられることなどが挙げられ，開発者側の条件としては，客観的に効果が確認されていない手法，試験的な手法など完成度に依存しないモジュールが収集できることが挙げられる．

これらを踏まえて，チャレンジ 1 に向けた課題として以下のものが挙げられる．

a)ユーザの卑近なニーズに応えられること（レポートやメールなど自分の文章や，口コミ，ブログなど他人の文章をチェックできる）
b)ユーザの多様なニーズに応えられること（モジュールの種類が豊富で充実している）
c)ユーザが使用する際の手間が少ないこと（簡易で直感的な操作方法）
d)ユーザの興味を引けること（直感的な面白さを備え，好印象の噂が広がるようにする）
e)モジュール作成のしきいが低いこと（仕様の理解と作成が容易）
f)モジュール作成のための支援環境が整えられること（既存モジュールの拡張や再利用が容易）
g)モジュール公開のしきいが低いこと（バグの不安やメンテの責任，クオリティの低さに対する懸念など精神面での障壁の排除）

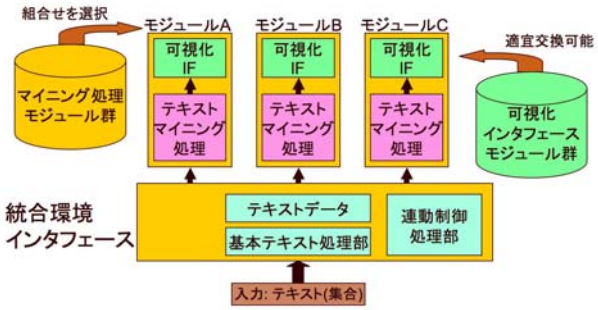

図 2 統合環境構成図

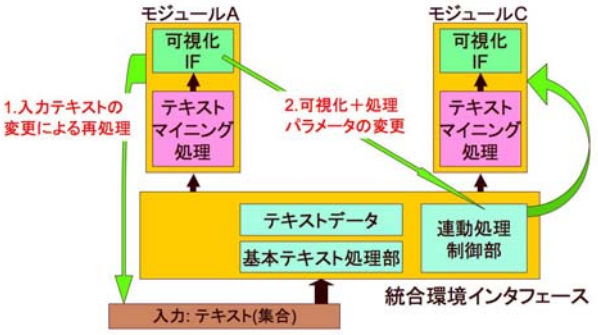

図 3 モジュール間相互インタラクション

h)モジュールの積極的な利用が見込まれること（多くの潜在的なユーザがいること）

§2 チャレンジ２：モジュール間での相互インタラクションの実現

図 2 の統合環境構成図に示すように，複数のテキストマイニング処理モジュールによる出力結果を，複数の可視化インタフェースモジュール上に同時に表示して，並列に並べて比較できるようにする．

特に可視化インタフェースモジュールは，出力インタフェースになると同時に，自モジュールや，他のモジュールへの入力インタフェースとしても機能させる（図 3）．単純に，入力テキストの変更により，各モジュールで再度処理を行った結果を表示させる機能（図 3 の 1.）に加えて，ある可視化インタフェースモジュールにおいて,「マウスによって選択されているデータが，他のモジュールのどこに出力されているかを，自動的に明示」したり,「他のテキストマイニング処理の実行と結果の表示」を相互に可能にする（図 3 の 2.）．

すなわち，チャレンジ２に向けた課題として以下のものが挙げられる．

a)他のモジュールで選択されているデータに対応するデータを捉えるために，モジュール間で共通のデータ構造に基づく連動の仕組みを用意すること

b)他のモジュールを操作するために，他に存在するモジュールの情報を共有するためのデータ構造を用意すること

§3 チャレンジ３：知識創発のための基盤環境の構築

複数のモジュールを併用して，試行錯誤による結果の分析を行える環境の中で，当たり前の結果だけではなく，頻度が低くても価値の高いデータ，パターンや知識を発見できる環境を構築する．

すなわち，チャレンジ３に向けた課題として以下のものが挙げられる．

a)精度や信頼度に依存しない多様なテキストマイニングモジュールが集められること

b)さまざまな角度からの分析が行える多様な可視化インタフェースモジュールが集められること

c)可視化インタフェースモジュールにおいて，直感的にデータ間の関係を捉えられ，分析作業に没入できること

この a) と b) は主にチャレンジ１の，c) は主にチャレンジ２の達成が必要条件となり，それぞれ技術的な側面もさることながら，ユーザや開発者の精神面においても，利用や開発をサポートする環境づくりが望まれる．

知識創発の達成のためには，環境を利用する際のユーザの思考過程と操作内容との関係を詳細に探る必要があるが，複数のモジュールを同時に用いて分析を進められる環境での，効果的なモジュールの選択方法，切り替え方や見せ方，効果的な操作方法など，利用方法に関する多くの点は明らかになっておらず，そのすべてを本チャレンジで実現できるかは定かではない．しかし，多様なモジュールが集められることを前提として，本環境に限らない知識創発を目指すシステム全般において，効果的な知識創発を実現する際の足がかりや雛形となりえる，基盤環境を構築していく．

2.2 統合環境の構成要素

構築する環境として，例えば，二次元ディスプレイ上の横 640pixel，縦 900pixel の領域を一つのモジュールのための領域として，この領域を縦と横に任意の数だけ並べることができるウインドウを作成する．現在の標準的なディスプレイであれば，1280 × 960pixel の解像度を出すことができるため，１画面内に２つの領域を表示できる．また 30 インチディスプレイ（解像度 2560 × 1600pixel）であれば，横に４つの領域 (図 4) を表示することができる．

現段階で，モジュールのフォーマットは，雛形クラスを作成した上で，それを継承したクラスを作成することを想定している．また各モジュールは，図 2 の基本的処理結果となる，入力テキストの基本的な情報（テキストデータ）を，共有して利用することを想定している．その他，オンラインでのリモートサービスによって，一部の処理をリモート環境で実行することや，処理の一部を並列分散化させることは，現在は想定していないが，検討課題の一つとして考えている．

複数のモジュールを１つのウインドウ内で動作させる環境の構成要素として，以下のものが挙げられる．

- 統合環境全体のウインドウシステム
- 環境が管理するテキストデータのデータ構造
- 環境を構成するモジュール群
- 環境とモジュールをつなぐインタフェース
- モジュールとモジュールをつなぐ（モジュール間の連動を可能にする）インタフェース

TETDMでは，今後これらに関する仕様を定めるなど，次節の計画にもとづいて環境の構築を進める．

2·3 TETDMチャレンジの計画

TETDMチャレンジの5年間の計画を以下に示す．

- １年目：統合環境の仕様の策定
- ２年目：モジュールの基本仕様の策定
- ３年目：モジュール間インタラクションの仕様の策定，ダウンロードサイトの立ち上げ
- ４年目：モジュール開発者支援
- ５年目：知識創発に向けた利用者支援

以下で，各計画の詳細について述べる．

§1 １年目：統合環境の仕様の策定

図2の「統合環境インタフェース」部分の仕様を策定する．すなわち，利用可能なモジュール群の中から，選択的にモジュールを選んで利用できる枠組み，また，入力されたテキストデータを保持するためのデータ構造を定める．データ構造は，多くのモジュールが利用する可能性の高いデータ（テキスト内の単語の出現情報，品詞情報や頻度情報など）についての情報ももつことを想定しており，図2の「基本テキスト処理部」において，それらのデータを作成する．これは主に，チャレンジ１のc) に関係する．

§2 ２年目：モジュールの基本仕様の策定

モジュール間のインタラクション部分を除いた仕様を決定する．各モジュールの入出力に関わる仕様，ならびに複数のモジュールを並列に動作させるにあたって，統合環境とモジュール間でデータの受け渡しを行うためのインタフェース（データ構造や関数）を定義する．この仕様の策定は，チャレンジ１のa)，b)，e) と主に関わっており，ユーザのニーズや開発のしやすさに応じた仕様を策定する必要がある．

なお，環境及びモジュールを作成するプログラミング言語は，普及率が高く汎用的なオブジェクト指向言語としてJava言語を用いることを想定している．また，Windows, Mac, LinuxといったOSの違いや，日本語の文字コードの違い，英語や数値データへの適用可能性，などを踏まえた仕様を検討する．

§3 ３年目：モジュール間インタラクションの仕様の策定

統合環境内で利用される各モジュールは，それぞれが独立に動作するだけでなく，あるモジュール内での操作が，他のモジュールにも反映される仕組みを導入する（チャレンジ２）．そのため，各モジュールが他のモジュールにアクセスするための方法と枠組みを策定する．この仕組みが実現されることは，チャレンジ１のd) や，チャレンジ３のb) とc) にも関わる．

§4 ３年目：ダウンロードサイトの立ち上げ

統合環境と各種モジュールをダウンロードできるサイトを立ち上げる．Webサーバに環境とモジュールのアプリケーションを置き，ダウンロード環境を構築する．また研究者が，各自が作成したモジュールをアップロードできるCGIを実装する．これは，チャレンジ１のh) や，チャレンジ３のa)，b) に関わる．

またこのようなサイトが必要な理由として，本環境は必ずしも万能ではなく，特に信頼度に依存しないモジュールの収集により，日々更新や改訂が行われること，また目標の1000以上の多数のモジュールのすべてを手元にダウンロードしておいておくことは，現実的ではないことが挙げられる．

§5 ４年目：モジュール開発者支援

モジュール開発者支援として，チャレンジ１のf)，g)，ならびにチャレンジ３のa)，b) に関わる支援の枠組みを検討する．すなわち，モジュールを容易に作成できるように，サンプルモジュールを充実させたり，既存のモジュールの再利用を促す環境を構築する．また，モジュールの公開に当たって，開発者がその内容についてのクレームや，保守の責任などの懸念が発生しないような枠組み，（例えば匿名の開発者を許すことなど）を検討し，積極的なモジュールの公開を促す．モジュールの信頼度を担保するため，また開発支援として，言語資源として正解データが利用できるモジュールについては，その評価支援を行うことも検討課題の一つとして考えている．

§6 ５年目：知識創発に向けた利用者支援

知識創発に向けた利用者支援として，チャレンジ１のa)，b)，d)，ならびに，チャレンジ３のc) に関わる支援を行う．すなわち，情報推薦の既存技術を応用するなどにより，ユーザの多様なニーズを満たすモジュールをスムーズに選択できる枠組みの検討（3·4節でも後述），ならびに，効果的な使用例と使用結果などのサンプルを積極的に公開する．

3. テキストデータ分析のための統合環境

本章では，テキストデータを分析するための，TETDMチャレンジで構築を目指す統合環境の概要について述べる．

3·1 統合環境の利用目的

テキストデータを分析する局面において，以下のような統合環境の利用目的が挙げられる．

1.特定のテキストデータの内容を詳細に理解したい
2.客観的にテキストデータの内容を把握したい
3.より多くの知見をデータから獲得したい
4.創造的活動において新たな行動戦略を創り出したい

図 4 サンプル環境画面 (解像度 2560 × 1000)

これらの目的は，テキストマイニングの各システムの目的としても，挙げられる機会が多い．しかし，単一のシステムで達成できることは限られており，これらの目的を，十分に納得がいくレベルで達成するためには，複数システムの出力を組み合わせていくことが肝要と考えられる．

3·2 統合環境の特徴

統合環境では，目的に応じたモジュールを選択でき，ユーザ独自の組合せにより処理を行うことができる．また，複数のモジュール間の入出力を連動させる環境により，ユーザの操作の手間を省き，データ認識の効率を高められる．特に各モジュールがもつ視覚化インタフェースを，出力のためだけではなく，入力のためのインタフェースとしても用いられるようにすることで，各モジュール上の出力を直感的に操作することによって，並列に表示されている他のモジュールの結果も連動して変更されるようにする．

すなわち，統合環境を用いないで，複数のシステム間でデータのやり取りを行う状況に比べると，以下を本環境の特徴として挙げることができる．

- ユーザは目的に応じて，容易に独自のモジュールの組合せを設定して，使用することができる
- システムの入出力操作の回数を減らすことができる
- 複数のシステムからの出力を対応づけて眺められる
- あるシステムの視覚化インタフェース上で，複数のシステムへの入力を，一度に直感的に与えられる

これら単純で直感的な操作と，データ認識の効率化により，テキストデータの集中的な分析と，データからのアイデア発生を促す．

3·3 サンプル環境

本節では，試験的に実装したサンプル環境について述べる*1．モジュール群は主に，次の 2 つの目的によって大別される．

1.特定のテキストについて，詳しい情報を知りたい
2.複数のテキスト間の関係や情報を知りたい

入力となるテキストデータについて，１つのテキストを入力する際には，その文の区切りを句点で，段落の区切りを特定のタグにより認識できる形式で入力する．複数のテキストを入力する際には，各テキスト内の文の区切りが句点で認識できる形式，また複数テキストを１つのテキストとして連結し，テキスト間の区切りが特定できるタグを挿入した上で入力する．

複数テキストを入力する際に１つのテキストとして連結するのは，ファイル入出力の回数を減らし実行時間の削減を図るとともに，テキスト間の区切りを，１つのテキストを入力とした際の段落の区切りと同等に見なすことで，１つのテキストを対象としたモジュールにも適用可能とするための措置となっている．

単一のテキストを入力として，そのテキストの情報を表示するサンプル環境上のモジュールを以下に示す．

1)テキスト表示（兼エディタ）
2)単語の頻度情報の表示
3)キーワード表示
4)要約表示 [相良 07]
5)テキストの一貫性表示 [砂山 08a]
6)意見文表示 [砂山 10]
7)主題関連部分の表示 [西原 09]

複数のテキストを入力とし，そのテキスト間の関係を表示するサンプル環境上のモジュールを以下に示す．

8)2 つのテキスト間の差分表示
9)クラスタリング結果表示 [Newman 04]
10)独自性表示 [砂山 08b]

*1 仕様策定のための実装であるため，最終的に構築される統合環境が本節で述べる形式を踏襲するとは限らない．

図4に，パネル（各モジュールの表示領域）を4つ並べたサンプル環境の画面を示す（左から順にモジュール7)，5)，9)，2)）．各モジュールは640 × 900pixelの大きさのパネル上に表示することができ，横に並べるパネル数は，実行時引数により変えることができる．

ユーザは，使用したいモジュールを，環境下部のボタンを押すことで選択できる．モジュール間の連動は一部でのみ実装されており，たとえば単一のテキストを入力した際に，3) のキーワード表示モジュールにおいてキーワードを選択すると，4) で選択されたキーワードを主題とした要約，5) で選択されたキーワードを主題とした一貫性表示を行うことができる．

また，複数テキストとしてレポート集合を入力として与えたときに，10) による独自性の表示に加えて，3つのテキスト表示パネルに，類似する2つのレポートを 1) で，それらの差分を 8) 上に連動させて表示することができる．代わりに 10) と 6) のモジュールを連動させて，意見文が書かれているレポートとその独自性を比較することもできる．

3·4 モジュール選択の指針

TETDMによって多くのモジュールが集められた際に，どのモジュールを使うべきかという問題の発生が懸念される．より具体的には，直面している問題に対する適切なモジュールの組合せが分からない場合と，使用するモジュールの出力結果がどの程度信頼できるかわからないという問題が起こると想定される．

前者の問題点に対しては，想定されるタスクに対して，推奨するモジュールの組合せを提示する機能を環境にもたせることで，解決できると考えている．推奨するモジュールの決定には，過去に使用したユーザのレビューなどをもとに，タスクと使用したモジュールの組合せに関するデータを収集することで，タスクごとに推奨されるモジュールの組合せデータを作成できると考えている．

後者の問題点に対しては，モジュールのダウンロードサイトにおいて，ユーザによるレビュー機能を設け，ダウンロード回数やユーザによる評価結果など，信頼度の指標になる値を表示する．また各モジュールについて，開発者側のコメントを載せるとともに，学会，論文誌などでの発表があれば，その出典の明記を促し，積極的な信頼性を担保する記述を増すことで，信頼度の評価ができるようにしていきたいと考えている．

4. TETDMチャレンジの位置づけ

本章では，関連する技術との比較，コミュニティの形成の視点からTETDMの位置づけを明確にする．

4·1 統合型データマイニングソフトウェア

与えられたデータから知識を得るには，データの前処理，本処理，後処理が必要となる．一般的な原著論文に記述されるデータマイニング手法は「本処理」の部分に主眼が置かれているが，前処理，後処理もデータマイニングにとって重要な役割を持っている [Fayyad 96]．本節では，前処理，本処理，後処理も併せて行うべく，複数の機能を備えて開発されたデータマイニングソフトを紹介した上で，本チャレンジの位置づけを行う．

有名な統計処理のフリーのソフトウェアとして R[R-Project] がある．R は前処理，本処理，後処理の組合せをユーザ自身が選び，統計計算処理（例えば，線形／非線形解析，時系列解析，クラスタリング）と可視化処理が行え，一つのデータを様々な手法によってマイニングすることが可能となっている．R はデータマイニングに慣れているユーザにとっては強力なソフトウェアとなる一方で，データマイニングに不慣れなユーザにとっては使いづらい可能性がある．TETDMでは，視覚的な入力インタフェース上で，直感的に動作可能な環境を構築し，データマイニングに不慣れなユーザにも簡便に利用可能な環境の構築を目指す．

Weka[Weka] や orange[orange] もフリーのソフトウェアとして，グラフィカルユーザインタフェースを備えており，データマイニングに不慣れなユーザでも利用しやすい．Weka はデータの機械学習を行うことができ，頻出するデータのパターンを発見することに優れている．TETDMでは，頻出パターンに加えて，希少価値があるデータのパターンを発見することも可能な環境を目指している．

前処理，本処理，後処理の組合せをユーザ自身が選ぶ必要がないソフトウェアには，ある目的に特化された機能を備えていることが多く，例えば，顧客の声を分析する目的のために開発されたソフトウェアとして，DIAMining, Text Mining Studio，TRUE TELLER，Text Mining for Clementine などが挙げられる [DIAMining, Mining Studio, TELLER, Clementine]．これらのソフトウェアはいずれも商用のソフトウェアとして，確立された技術をもとにした信頼度の高いマイニング結果を出力する．しかし信頼度が高い結果は，既知の一般的知識を多く含み，発見的な知識を得ることは難しい．また有償のソフトウェアであるため，誰もが気軽に用いることはできない．TETDMでは，必ずしも信頼度が高くないが何らかの特徴があり，応用の可能性を秘めたデータを，多様なモジュールの組合せとデータとのインタラクションを通じて発見できる環境の構築を目指す．

4·2 データマイニングの統合環境

本節では，視覚的なデータマイニング技術の統合環境として，今までに開発されて来た環境やプロジェクトについて述べる．

VidaMine[Kimani 03] は知識発見プロセスの前処理，本処理，後処理の全てを単一の環境で行うことを目的として開発された．しかし対象はテキストではなく，データマイニングのためのモジュールも多くは集められていない．

辞書や機械翻訳などの言語資源を言語サービスとして登録し，共有可能にするインターネット上の多言語サービス基盤として，言語グリッド [言語グリッド] が挙げられる．さまざまなサービスを登録して相互に利用できる環境を目指す点は同じであるが，範囲が言語サービスに限られており，サービスという側面から主に精度が重視される．本チャレンジでは，テキストの分析結果をもとにユーザの試行錯誤を促し，新たな発見や発想を促す環境の構築を目指す点が異なる．

テキストマイニングのための統合的ツールに GATE [GATE] があり，既存のツールの再利用を目指している．研究の実用化のためのベンチマークテストデータの設定など，研究の再利用を意識した作りで，専属の開発チームによるリリースが行われているが，精度が重視されるモジュール構成で，多様なモジュールを集められる枠組みにはなっていない．

ユーザとシステム間のインタラクションを想定した統合環境に LanguageWare [Language] がある．入力テキストの言語の推定，単語の抽出，品詞の推定，単語の正規化，固有表現抽出などのテキスト分析を行うことができる．また，UIMA（Unstructured Information Management Architecture）[Ferrucci 04] と呼ばれる基準に基づいてコンポーネントを作成しているため，この基準に基づく他のコンポーネントとの組合せも可能となっている．LanguageWare は，主にビジネスの現場にいる人をユーザとして想定しており，利用にはテキストマイニングの知識や経験が求められる．本チャレンジでは，学生や主婦など，PC は利用するがテキストマイニングという言葉を知らないユーザも想定しており，単純かつ直感的に用いられ，利用価値がある環境の構築を目指す．

種々の言語のデータを，さまざまなモジュールで扱えるようにするためのミドルウェアアーキテクチャとして，Heart of Gold [Heart] がある．コーパスへの自動かつ多次元のアノテーション，XML ベースでのモジュール同士の結合などを行える．これは主に，言語データの中から共通パターンを見つけ出すことなどを念頭においているが，本チャレンジでは希少価値のあるパターンを発見できる環境づくりを目指す．

U-Compare[狩野 08] は処理の実行順序がプログラマブルで，相互依存するコンポーネント（モジュール）を入出力定義により自動的に組み合わせられる．また，UIMA にも準拠している．U-Compare においても複数コンポーネントの結果を視覚化して同時に並列表示するが，本研究では，複数の可視化モジュール間でのインタラクションを可能にし，ユーザの集中的な試行錯誤を促せる環境の構築を目指す．

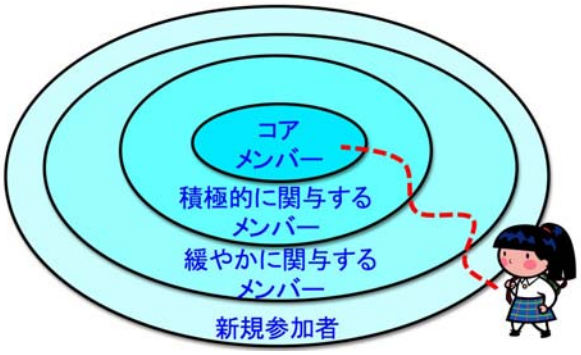

図 5 正統的周辺参加

また入出力やデータの取り扱いに関して，広く用いられている UIMA フォーマットの利用に関して，最終的にどのようなデータフォーマットを利用するかについて，現在検討中の段階であり，UIMA に準拠させることや，UIMA への適用可能性を残すことも一つの選択肢として考えている．しかし幅広いモジュールの開発と収集のためには，大学等で一通りプログラムを学習した段階の，開発経験が少ない人でも理解が容易なフォーマットが望まれるため，最低限の内容を備えたフォーマットについて，まずは検討を進めていきたいと考えている．

これら既存のテキストマイニングシステムにおいては，信頼度の高いモジュールが多く，ヒューリスティックな手法（発見的手法）や思いつきによるアドホックな手法に基づくモジュールが比較的少ない．客観的な評価を得たシステムをモジュールとして収集することはもちろん望ましいが，モジュール作成のしきいも大きく上がるため，モジュール収集の範囲が狭まったり，新技術の速報性にも欠ける可能性がある．本チャレンジでは，信頼性のしきい値を下げることによって，多様な技術のモジュールを集められる枠組みを構築し，ユーザの思考を活性化し，新たな知識の発見を促せる環境づくりを目指す．

また既存のシステムを，本チャレンジの環境で外部処理として呼び出す事は可能と考えており，それら既存のシステムや，既存の言語資源等を可能な範囲で統合環境に組み入れることは検討課題のひとつとして挙げられる．

4·3 コミュニティの形成

本節ではコミュニティの形成という立場から TETDM の位置づけや意義を明確化する．

一般に，コミュニティのメンバーは皆が同じ立場で関わるわけではなく，(1) 一人ないし少数のコア・メンバー，(2) 積極的な参画や寄与を行うメンバー，(3) 緩やかにそのコミュニティに関与するメンバー，と複数の異なる立場で関わっていく（図 5）．すでに形成されたコミュニティに新たに参画するメンバーは，(図 5）の周縁部から参画し，興味が合えば徐々に中心へと，段階を経ながら移行していく．これは正統的周辺参加（Legitimated Peripheral

Participation) [Lave 91] と呼ばれる参加形態で，このような形態が適切に維持されることが，コミュニティの活性化・発展に繋がる．TETDM が目指すコミュニティ形成もこの形態に類すると考えており，ユーザ，研究者の両方において，メンバーとしての参画を促していける形態を模索している．具体的には，学会や研究会において，幅広くユーザやモジュールの開発者を募集することで，上記 (3) に該当するメンバーを集め，TETDM と連動する研究会を開催することで，上記 (2) に該当するメンバーが集まれる場所を用意する．

この正統的周辺参加の観点の下で，複数の研究主体が協力，協調しつつ研究の発展を図るための求心力として，以下の 4 つが挙げられる．

§1 情報の提供

学会誌の解説記事や書籍，あるいはブログや WIKI などのネット媒体を用いて，関連する手法や技術の横断的かつ網羅的な情報ポータルを構築する試みがある．例えば，テキストマイニング・学習に関するものとしては，人工知能学会誌の「私のブックマーク」[長谷川 01] や朱鷺の杜 Wiki [Wiki] などがそれに当たる．

§2 標準インタフェースの策定

同種の，あるいは類似した目的のシステム間で，扱えるデータや知識の共通化を図ることで，それらの比較や検証を容易にしたり，システムの入出力のプロトコルやフォーマットを共通化することで，複数のシステムがプリプロセス/ポストプロセスという形で連携できるようにしたりする試みがある．例えば，柔軟な形態素解析システム間接続を行なうためのドライバモデルである MACD [MACD, 松田 99] などがそれに当たる．これによって，システムの可搬性や可換性の向上が期待できるため，システム開発者にとっての動機づけになり得る．近年では，Yahoo! API[yahoo] など API (Application Program Interface) を規定することで，利用者に WEB サービスの機能を部分的に開放するような試みも増えている．

§3 テストコレクション・共通課題の提供

システム間の比較や性能評価のために共通課題や評価データを提供し，競争的な環境を作る試みがある．特定のテキストコーパスなどを対象とするようなクローズドな課題であれば，主催者が人手をかけて事前に作成した正解データセットや標準回答などを用意し，それに基づいて参加者が提出した結果を評価する方法がとられる．例えば「動向情報の要約と可視化」ワークショップなどがそれに当たる．一方，Web や複数年にわたる新聞記事のように，対象が莫大な，もしくはオープンな場合は，一部を抜き出したテストコレクションや課題を提供し，各参加者がシステムの実行結果を持ち寄ってその集合を解空間の全体集合と見做した上で，各々のシステムを評価する方式（プーリング方式）や，実行結果を主催者が人手で評価して順位付けする方式などがとられる．例えば，前者としては TREC [TREC] などが，後者としては InfoVis Contest [Plaisant 07] などがそれに当たる．

§4 システム・ツールの提供

直接的な研究成果であるシステムの提供や，データの整形や整合性チェックのツールの提供を通じて，他者が開発したシステムとの比較を可能にしたり，それらのシステムやツールを利用，拡張して別のシステムを開発できるようにする試みがある．例えば，近未来チャレンジ「情報編纂の基盤技術」における可視化プラットフォーム [松下 09] などがそれに当たる．

4·4 コミュニティ形成への道筋

研究主体のコミュニティには，類似した研究テーマの実践を行う主体が情報交換を目的として集う形態（Community of Practice，以下 CoP)[Lave 91] と，ある目的の下で，異なる専門性や関心を持った主体が連携や協同を目的として集う形態（Community of Interests，以下 CoI)[Arias 00] の，二つの形態が考えられる．特に後者は，対象とする課題の複雑化や分野横断化に伴い，近年様々なかたちでの協同が模索されている．TETDM は，この二つのコミュニティの側面を併せ持つプロジェクトとして位置づけられる．すなわち，4·3·1 節のような情報提供のためのポータルをベースとして，多くのテキストマイニング技術を統一的に連動させて扱えるようにすることで，4·3·4 節に相当する環境を提供する，いわば「ワンストップサービス」としての場の実現を目指している．そのため，この場は，CoP と CoI が邂逅する場ともなり得る．

すなわち，テキストマイニングツールを開発する人々 (Community of Practice) からすれば，各々が作成したツールの比較競争の場とみることができ，それらを利用する人々 (Community of Interests) にとっては，提供されたマイニングツールを利用して自らの問題の解決や解消に役立てられる場となる．

CoP の立場からすれば，自らのシステムを利用するユーザを確保できる点や，様々な観点からの競争や連携が期待できる点が，システム提供の動機づけとなる．また，CoI の立場からすれば，ニーズに応じた技術の選択や比較が容易なため本来の興味である課題（分析作業）に集中できる点が，利用の動機づけになる．

これらに加え，コミュニティに参画するための最初のしきいを可能な限り下げることで，より大きなコミュニティの形成を目指す．すなわち，TETDM が提供する環境について，CoP の開発者の立場であれば何らかのモジュールを作成することが，CoI のユーザの立場であれば本環境を用いてみることが，コミュニティ参画へきっかけになると考えられる．したがって，開発のためのモジュールの仕様を厳しく定めないとともに，モジュールの雛形やサンプルを提供することで「自分にも作れそう」と思える仕様，またユーザとして使用する際の手間，および使用の説明を最小限に留められ，直感的に「使いやすそう」と思える，またその出力が魅力的で「面白そう」と

思える見た目を用意する．

5. 社会の発展や研究促進に向けた応用事例

本章では，TETDM チャレンジによる環境が，社会や研究開発の場面で用いられる応用事例について述べる．

5·1 社会や研究開発における卑近な使用例

現在の世の中は，多くの情報を獲得するとともに，それらをいかに分析して次の行動につなげていくかが問われている．その際に情報を多角的に分析できるツールは必須と考えられる．コンピュータを使っていて電子テキストを扱わない人はおらず，簡便で実用的な環境の上で，多角的にテキストを分析できるツールへのニーズと期待は高いと考えられる．多くのユーザの利用が見込まれる卑近な使用例としては，メールの作成支援として，自分の作成したメールの誤字や脱字，敬語のチェック，読んだ人が受ける感情の推定結果などを表示することや，Web ページ検索の結果のテキスト集合や，興味あるブログやつぶやき [twitter] のテキスト集合から，キーワードや関連語情報，分類結果と各分類の要約を提供することなどによる，情報へのアクセス支援と，新たな興味への発想の手がかりを提供することが考えられる．

また，研究開発者が作成したシステムの評価を行う際には，類似システムとの比較が必要な場合が多い．新しい技術が研究論文として発表されても，それが実際に活用されるためには，論文の著者からツールをもらうか，独力で実装する必要がある．１つの共通の環境が存在して，そのモジュールとしてダウンロードが可能になれば，既存研究との比較も容易になると考えられる．

5·2 大局視と局所視の融合による探索的データ分析

動向情報の分析など探索的データ分析が必要な場面では，データの全体像（大局視，overview）と各データの詳細（局所視，detail）の両方を提示することが重要というコンセプトがある．しかし多くの可視化モジュールにおいては，この大局視あるいは局所視のどちらかに適している（特化している）．動向情報には，時間的動向情報と空間的動向の二種類があり，それぞれに適した可視化表現が異なる（例えば，時間的動向は折れ線グラフ，空間的動向は地図など）ので，それぞれに適したモジュールを複数用意する必要がある．また，時間的／空間的両方の性質を持つ動向情報の存在や，探索的分析においては時間的，空間的両面から分析を進める必要があるため，モジュール間の連携が不可欠となる．

ユーザの観点からの利点として，大局視，局所視（あるいは空間的動向，時空間的動向）それぞれについて多様な可視化技術が存在するため，その中からユーザが好むものを選択して分析環境を構築できることが挙げられる．また開発者の観点からの利点として，例えば，大局視用の新しい可視化技術を開発し，他の大局的可視化手法と比較したい場合，局所視部分などの他の要素については条件を揃えて実験をしたい，というケースはよくある．そのような要望を，TETDM の環境では容易に叶えることができる．

5·3 電子カルテのテキストマイニング

病院に蓄積されているデータのほとんどは構造化されていないテキストデータであると言われており，テキストデータから興味深く重要なパターンを抽出し，その分析結果を視覚化するテキストマイニングが注目を集めている．特にデータ量が膨大になると，データ自体を整理することさえ困難な現状がある．そこで，このような大量のデータから医療行為に活用できる知識を獲得する際に，複数のテキストマイニング技術を連動させる環境が活用できる．

例えば，多くの電子カルテのテキストデータ集合をその内容に応じてクラスタリングする際に,「テキストデータ集合からキーワードを抽出するモジュール」と連動させて「クラスタリングモジュール」の分類基準を決定することで，より柔軟なクラスタリング結果を表示できる．また，クラスタを選択した際には，各クラスタの代表的なカルテの要約を表示するための「テキスト自動要約モジュール」や，ピックアップしたクラスタ内のカルテ間の関係を示すために「ネットワーク可視化モジュール」を利用することなどにより，隠れた関係性や意味を見つけ出すことができる．

蓄積された電子カルテのテキストデータ（看護記録・経過記録）をテキストマイニングにより解析し活用すれば，医師，看護師等の教育や評価に活用できるほか，ベテランの医師，看護師でなくても適正な記録がなされているかどうかの判断材料とすることで，医師，看護師でなくても診療情報管理士による質的監査が可能になると考えられる [Kushima 10]．他にも，患者の疾患の予後・予後因子・治療成績・医療技術の安全性等を評価する臨床研究データを集積し，臨床研究のコスト削減・効率化・品質向上に繋げるなど，テキストマイニング技術の応用の幅は広く，これら多様な目的のために，個々にシステムを構築することは現実的ではない．

5·4 開発済みモジュールの適用範囲の拡大

研究者の立場の応用として，既存ツールの新しいドメインへの適用可能性 [Daume 06] を探ることへの応用が考えられる．例えば言語処理の分野で「言語解析ツール（形態素解析器，固有名詞抽出器，係り受け解析器，評判分析器など）」をそれらを学習させたドメイン（例えば新聞記事，Wall Street Journal コーパスなど）と全く異なるドメイン（例えばブログ，twitter など）で適用する場合，どのようなところで問題が起こるか，どのようなエラーを引き起こすか，そしてどうすれば新しいドメイン

へ適応できるか (domain adaptation) という研究が盛んに行われている．今後更に多様なドメインが増えていくと予想されるため，このようなニーズは今後増していくと考えられる．

そのような場合,「ドメイン適用性を調べたいツール（モジュール X）」と「2 つのテキストを比較して，共通点や相違点を列挙するモジュール」および「(列挙された) データを分類して視覚的にクラスタリングするモジュール」と連動させることで，ある新しいドメインに対するモジュール X による処理結果（テキスト A）と，他に用意した理想的な処理結果（テキスト B）とを比較して，視覚的に分析を行うことが容易になる．また，他に競合しそうなツールによる結果を並べて見たり，類似ツールとの結果の比較分析も可能になる．

6. 結　　論

TETDM チャレンジでは，複数のテキストマイニング技術を柔軟に組み合わせて使える環境を構築し，それらを広く提供することを目指している．本環境により，複数の技術を用いたいユーザの環境が整えられ，ニーズに応じたモジュールを選択した上で，集中的して作業を行うことができるようになると期待できる．

個別のテキストマイニング技術を開発する際に，他の技術との連携を意識して視野を広げつつ，本環境に統合することができれば，多くの研究が認知，実用化されるようになり，ユーザの情報の多角的な分析に基づく創造的開発，研究，経営戦略の立案などが支援されると期待できる．

TETDM チャレンジが目指す環境は，より多くの人に認知されて初めて成立するため，多くの方々の積極的なご助言とご助力を賜りたいと考えている．多くの人々が集団となったときのエネルギーがあれば，テキストマイニングの技術開発は飛躍的に進むと期待でき，またテキストマイニング分野での成功が認知されれば，他の技術開発分野でも同様の動きが現れ，世の中のさまざまな技術の格段の進歩が見込まれる．

謝　　辞

本論文の完成に当たり，査読者の方々から，大変有益なご助言を頂きました．ここに記して感謝致します．

◇ 参 考 文 献 ◇

[Arias 00] E. Arias, H. Eden, G. Fischer, A. Gorman, and E. Scharff: Transcending the Individual Human Mind: Creating Shared Understanding through Collaborative Design, ACM Trans. on Computer-Human Interaction, Vol.7 No.1, pp.84 – 113, (2000).
[Clementine] Text Mining for Clementine（http://www.spss.co.jp/software/modeler␣a/）
[Daume 06] Hal Daume III and Daniel Marcu: Domain Adaptation for Statistical Classifiers, Journal of Machine Learning Research, Vol 26, pp.101 – 126, (2006).
[DIAMining] DIAMining（http://www.mdis.co.jp/products/diamining/）
[Fayyad 96] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth: Knowledge Discovery and Data Mining: Towards a Unifying Framework, KDD, pp.82 – 88, (1996).
[GATE] GATE（http://gate.ac.uk/）
[言語グリッド] 言語グリッド（http://langrid.nict.go.jp/jp/）
[長谷川 01] 長谷川隆明：私のブックマーク「テキストマイニング」，人工知能学会誌, Vol.16. No.6, p.893, (2001).
[Heart] Heart of Gold（http://heartofgold.dfki.de/）
[狩野 08] 狩野芳伸，辻井 潤一：UIMA を基盤とする相互運用性の向上と自動組み合わせ比較ー国際共同プロジェクト U-Compare, 情報処理学会自然言語処理研究会報告, Vol.2008, No.67, pp. 37 – 42, (2008).
[Kimani 03] S. Kimani, S. Lodi, T. Catarci, G. Santucci and C. Sartori: VidaMine:A Visual Data Mining Environment, Journal of Visual Languages and Computing, Vol.15, No.1, pp.37 – 67, (2004).
[Kushima 10] M. Kushima, K. Araki, M. Suzuki, S. Araki, and T. Nikama: Graphic Visualization of the Co-occurrence Analysis Network of Lung Cancer in-patient nursing record, proc. of The International Conference on Information Science and Applications(ICISA 2010), pp.686 – 693, (2010).
[Language] LanguageWare（http://www.ibm.com/software/ jstart/languageware）
[Lave 91] J. Lave and E. Wenger: Situated Learning: Legitimate Peripheral Participation, Cambridge Univ. Press, (1991).
[MACD] MACD (http://chasen.aist-nara.ac.jp/macd/)
[松田 99] 松田寛: 形態素解析システム相互接続ドライバモデル MACD の設計,「言語資源の共有と再利用」シンポジウム論文集, (1999).
[松下 09] 松下光範，加藤恒昭：情報編纂研究促進のための試み, 人工知能学会誌, Vol.24, No.2, pp. 272 – 283, (2009).
[Mining Studio] Text Mining Studio（http://www.msi.co.jp/tmstudio/）
[Newman 04] Newman, M.E.J.: Fast Algorithm for Detecting Community Structure in Networks, Physical Review E 69, 066113, pp. 1 – 5, (2004).
[西原 09] 西原陽子，佐藤圭太，砂山渡：光と影を用いたテキストのテーマ関連度の可視化, 人工知能学会論文誌, Vol.24, No.6, pp.480 – 488, (2009).
[orange] orange (http://www.ailab.si/orange/)
[Plaisant 07] C. Plaisant, J. D. Fekete, and G. Grinstein: Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository, IEEE Trans. on Visualization and Computer Graphics, Vol. 14, No.1, pp.120 – 134, (2008).
[R-Project] R-Project（http://www.r-project.org/）
[相良 07] 相良直樹，砂山渡，谷内田正彦：サブトピックを考慮した重要文抽出による報知的要約生成, 電子情報通信学会論文誌, Vol.J90-D, No.2, pp.427 – 440, (2007).
[砂山 08a] 砂山渡：テキストの話の流れを視覚化するインタフェース, 第 22 回人工知能学会全国大会, 1B1-1, (2008).
[砂山 08b] 砂山渡，川口俊明：内容の独自性の視覚化によるレポートの独自性評価支援システム，人工知能学会論文誌，Vol.23, No.6, pp.392 – 401, (2008).
[砂山 10] 砂山渡，川口俊明，田村幸寛：レポートの課題との関連度と意見文抽出による情報量評価支援, 電子情報通信学会論文誌，Vol.J93-D, No.10, pp.2032 – 2041, (2010).
[TELLER] TRUE TELLER （http://www.trueteller.net/）
[TREC] TREC (http://trec.nist.gov/)
[twitter] twitter（http://twitter.com/）
[Ferrucci 04] Ferrucci, D. and Lally, A. : UIMA: an architectural approach to unstructured information processing in the corporate research environment, Natural Language Engineering, Vol.10, No.3-4, pp.327 – 348, (2004).
[Weka] Weka（http://www.cs.waikato.ac.nz/ml/weka/）
[Wiki] 朱鷺の杜 Wiki(http://ibisforest.org/)
[yahoo] Yahoo! API (http://developer.yahoo.co.jp/)

〔担当委員：阿部 明典〕

2010 年 12 月 28 日　受理

著　者　紹　介

砂山　　渡（正会員）

1995 年大阪大学基礎工学部制御工学科卒業．1997 年同大大学院博士前期課程修了．1999 年同大大学院博士後期課程中退．同年同大学院助手，2003 年広島市立大学助教授，2007 年同准教授，現在に至る．博士 (工学)．人間の創造活動を支援する研究に興味を持つ．電子情報通信学会，言語処理学会，IEEE, 各会員．

高間　康史（正会員）

1994 年 東京大学工学部電子工学科卒業．1999 年同大学院博士課程修了．1999–2002 東京工業大学大学院総合理工学研究科助手, 2002–2005 東京都立科学技術大学助教授, 2005 年より首都大学東京システムデザイン学部准教授．博士 (工学)．Web Intelligence や情報可視化，知的インタフェースの研究に従事．主要著書は「インテリジェントネットワークシステム入門」(コロナ社)．IEEE，日本知能情報ファジィ学会，情報処理学会，電子情報通信学会各会員．

Danushka Bollegala

2005 年東京大学工学部電子情報工学科卒業，2007 年同大学院情報理工学系研究科修士課程修了，2009 年同研究科博士課程修了（短縮修了）．博士（情報理工学）．現在，同研究科助教．

西原　陽子（正会員）

2003 年大阪大学基礎工学部卒業．2005 年同大大学院基礎工学研究科博士前期課程修了．2007 年同研究科博士後期課程修了．博士（工学）．日本学術振興会特別研究員を経て，2008 年東京大学大学院工学系研究科助教，2009 年同講師，現在に至る．コミュニケーション支援，サービスデザインに興味を持つ．情報処理学会，医療情報学会，機械学会，各会員．

徳永　秀和（正会員）

1986 年東京工業大学大学院総合理工学研究科修士課程修了．同年新日本製鐵（株）勤務．1993 年高松工業高等専門学校講師．2005 年同校助教授，2007 年同准教授，2009 香川高等専門学校准教授，現在に至る．博士 (工学)．Web からの知識獲得，知識共有の研究に興味を持つ．情報処理学会，日本知能情報ファジィ学会, 各会員．

串間　宗夫（正会員）

1987 年宮崎大学大学院工学研究科修士課程修了．2003 年同大大学院工学研究科博士後期課程修了．2008 年同大大学院医学系研究科博士課程医学専攻入学，現在に至る．博士 (工学)．地方公務員．医学系では，癌治療，診療情報，電子カルテ，地域医療連携，工学系では，MOS アナログ集積回路，多値論理回路，教育工学，に興味をもつ．日本医療情報学会，電子情報通信学会，多値論理研究会，バイオメディカル・ファジィ・システム学会，各会員．

松下　光範（正会員）

1995 年大阪大学大学院基礎工学研究科物理系専攻制御工学分野博士前期課程修了．同年日本電信電話 (株) 入社．2008 年 関西大学総合情報学部准教授，2010 年 同教授，現在に至る．自然言語理解，情報可視化，ヒューマンコンピュータインタラクションに関する研究に従事．情報処理学会，日本知能情報ファジィ学会，日本バーチャルリアリティ学会，ACM 各会員．博士 (工学)．