

統合環境 TETDM を用いた知識創発支援の枠組み

Framework of Knowledge Creation Support with TETDM

砂山 渡^{*1}
Wataru Sunayama

¹ 広島市立大学大学院 情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

Abstract: In TETDM 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. Therefore, TETDM is suggested as a system that can integrate numerous tools and can contribute to our creative work. In this paper, framework of knowledge creation support with TETDM is described. Knowledge creation consists of “trials and errors” and “logical interpretations of the facts”.

1 はじめに

人工知能学会全国大会の近未来チャレンジテーマとして進められている TETDM (Total Environment for Text Data Mining: テトディーエム)¹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、社会的創造的活動を支援できる環境としての提供を目指している [砂山 11, 砂山 13, 砂山 14]。近未来チャレンジでは、5 年以内の目標達成を必要条件としており、現在 4 年目 (2014 年の全国大会で 5 年目に入る) の TETDM は、統合環境の正式公開を 2015 年 4 月と目標を定めて進めている (図 1)。

これまでに、複数のモジュールを統合的に扱う環境の枠組み作りを行い、2011 年 12 月の試用 α 版の公開を皮切りに、2014 年 3 月現在でバージョン 0.55 までを公開している。統合環境ならびに必要な情報の公開は、TETDM サイト上で行っており、継続的にアクセスとダウンロードがなされている。

本稿では、この統合環境 TETDM を用いた知識創発支援の枠組みについて述べるとともに、現在までの実装の状況と今後の見通しについて述べる。本研究での知識創発は、データとツールを用いた創造的思考と考えられ、さまざまな可能性を列挙する発散的思考のために、多様な観点からデータを分析して知見を集める場面と、集められた知見の集合を、収束的思考により一つの汎用的な知識となるようにまとめる場面からなる。

(連絡先) 砂山渡, 731-3194, 広島市安佐南区大塚東 3-4-1, 広島市立大学大学院情報科学研究科, sunayama@hiroshima-cu.ac.jp

¹TETDM サイト: <http://tetdm.jp/>

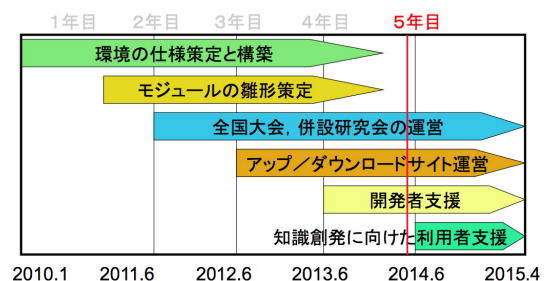


図 1: 近未来チャレンジ TETDM のスケジュール

2 TETDM 統合環境の構成

本章では、TETDM 統合環境の構成と、利用者支援のために新たに実装した機能について述べる。TETDM は、あるテキストを入力したときに、その分析処理を行う処理モジュールと、処理結果を可視化して出力する可視化モジュールを複数備えた環境となっている (図 2)。またこれらのモジュール群は、単独の開発者によって提供されるものではなく、任意の開発者がモジュールを作成して利用することができる。

TETDM は、独立した複数のパネル内に、処理モジュールと可視化モジュールを 1 つずつペアとしてセットすることで動作する。各パネルにセットされる異なる開発者によって作成されたモジュールが、それぞれ独立に動作するだけではなく、それらを協調的に連動させることができる点が TETDM の特徴となっている。

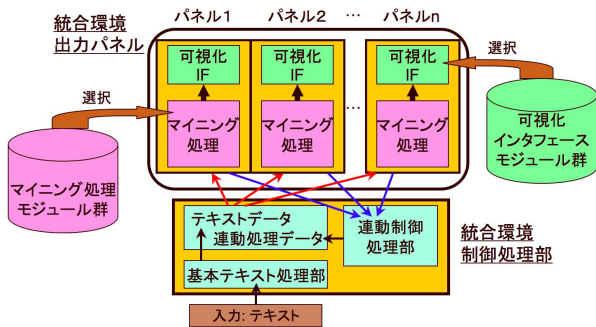


図 2: TETDM 統合環境構成図

以下で各部の説明を述べる²。

2.1 入力：テキスト

TETDMに入力されたテキストは「セグメント」「文」「単語」の3つに分割して扱われる。「単語」へ区切る際は、形態素解析器を用いて単語に分割する。この際、指定した品詞の単語だけを、キーワードとして取り扱う。「文」に区切る方法は、テキスト中の句点記号(「。」や「。」)をもとに分割する。「セグメント」に区切る方法は、特定の文字列をテキスト中に挿入し、その文字列をもとに分割する。その後、単語の出現情報や頻度情報の計算、キーワードやセグメント間の関連度計算を行った結果をデータ構造に格納し、このテキストデータをもとに各モジュールが処理を行う。

2.2 マイニング処理モジュール

マイニング処理モジュールは、統合環境内のテキストデータをもとに、テキストの理解や分析に役立つ情報をテキストから抽出する。またマイニングという言葉にこだわることなく、テキストに何らかの処理を施すモジュールも対象とする。現在までに30以上の処理モジュールが作成、公開されている。マイニング処理モジュールの処理結果は、可視化インタフェースモジュールに渡されて出力される。

2.3 出力：可視化インタフェースモジュール

可視化インタフェースモジュールは、マイニング処理モジュールによる出力結果を可視化する³。可視化インタフェースモジュールでは、入力として受け取れるデータ型(boolean, int, double, String型とその一次

²構成の詳細は[砂山 14]を参照していただきたい。

³マイニング処理モジュールの結果によらず、統合環境のテキストデータを可視化するモジュールであっても良い。

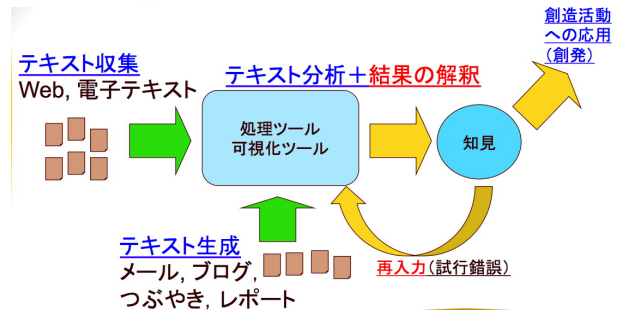


図 3: テキストマイニングのプロセスと知識創発

元配列、二次元配列(String型以外)の11種類)を定め⁴、そのデータの意味を表す整数型変数との組合せにより、マイニング処理モジュールからデータを受け取ることができる。現在までに30以上の可視化モジュールが作成、公開されている。

3 TETDMを用いた知識創発の枠組みと実装

本節では、統合環境 TETDM を用いた知識創発支援の枠組みについて述べる。

3.1 テキストマイニングのプロセスと知識創発

図3に、テキストマイニングのプロセスと知識創発との関係を示す。テキストマイニングのユーザは、テキストをツールに入力してその分析結果を取得する。ユーザは、得られた結果に対して解釈を与え、意味を理解することによって知見(一つの考え)を得る。また得られた知見について、その知見の信頼性や妥当性の確認しながら、質の高い知見を得るための試行錯誤を行う。そして集められた知見を総合的に解釈することで、実際に活用可能なレベルの知識へと変換させる(創発)。これらの用語は以下のようにまとめられる。

知識創発：知見を積み重ねることで汎用可能な一般化された新たな知識を発見すること

試行錯誤：得た知見について根拠や理屈を確認するため、または新たな知見を得るために、再入力を行うこと(発散的思考)

結果の解釈：テキストの意味や性質を理解して知見(一つの考え)を得ること(収束的思考)

⁴TETDMはJavaで実装されている。

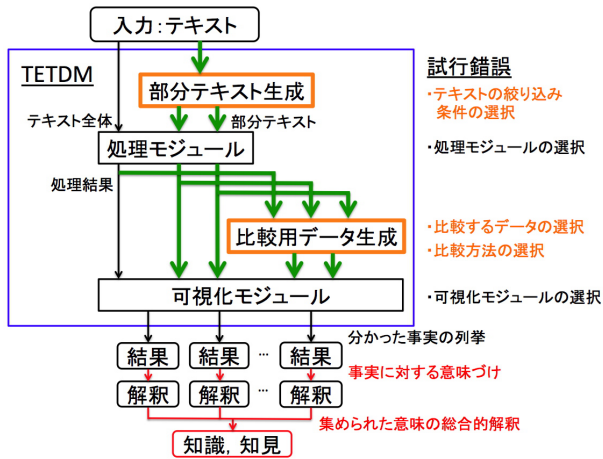


図 4: TETDM を用いた知識創発の枠組み

3.2 TETDM を用いた知識創発の枠組み

前節の定義に基づいて、知識創発を実現するためには、多くの知見を得るための試行錯誤を繰り返す必要があり、この試行錯誤のバリエーションを用意することで、得られる結果の幅が広がると想定される。

図4に、TETDMを用いた知識創発の枠組みを示す。2章で述べた基本的な構成のTETDMにおいては、使用する処理モジュールと可視化モジュールを選択すること（ならびに各モジュール内で用意された機能を用いること）で試行錯誤を行っていた（図4右の「処理モジュールの選択」「可視化モジュールの選択」）。新しい枠組みにおいては、テキストデータを絞り込む「部分テキスト生成」、ならびに全体と部分、あるいは部分同士を比較可能にする「比較用データ生成」の機能をTETDMにもたせることで、試行錯誤の幅を広げる。前者は、データ量が多い場合には必須となり、データ量が少ない場合でも、まずはデータの一部に着目して、そこからわかる事実を見つけることの重要性は高い。後者は、一部のデータの特徴を捉えるためには、基準となるデータ全体の傾向との比較や、他の部分データとの傾向を比較する必要がある。

また、結果の解釈については、通常のツールは結果を出力するのみとなっており、その結果の解釈を積極的に支援するものにはなっていない。出力された結果に意味づけを行うこと、また意味付けされた結果を集めて整理する環境を用意することが、知識創発につながると考えられる。

これらの「試行錯誤」と「結果の解釈」は、創造的思考における、発散的思考と収束的思考 [Guilford 59] に対応すると考えており、既存研究から得られている知見も適宜採り入れていきたいと考えている。現時点での枠組みの詳細について以下で述べる。

3.3 試行錯誤

知見を得るための試行錯誤として、以下の3つを想定している。

1. 想定内外のツールの選択と組合せ
2. 部分テキストの生成
3. 比較用データの生成

1. はTETDMにおける使用するツールの選択における試行錯誤、2. はテキストデータの絞り込み方法における試行錯誤、3. は絞り込んだテキストを解釈するための試行錯誤となる。この試行錯誤の過程は、創造的思考における発散的思考に位置すると考えられる。

3.3.1 想定内外のツールの選択と組合せ

TETDMにおいては、使用する各パネルにおいて、処理ツールと可視化ツールを1つずつセットして利用する。そのため、各パネルにおける処理ツールと可視化ツールの組合せ、並びに、複数のパネルで用いるツールの組合せには自由度がある。

ここで、さまざまなツールを用いることで、多様な分析結果を獲得できると考えられ、用いるツールの組合せによって、想定されるツールの組合せから得られる知識、ならびに想定外のツールの組合せから得られる知識とに分けられる。

想定されるツールの組合せとは、「ユーザが予め指定するツールの組合せによる分析」または「ツールの作成者が予め指定するツールの組合せによる分析」を指す。すなわち、ユーザまたはツールの作成者があらかじめ利用をイメージしたツールの組合せによってツールを利用して、結果を獲得する場合となる。この場合、利用方法の見通しが立っているため、意味のある結果を得やすい反面、得られる結果が予想の範囲内にとどまることも多いと考えられる。

想定外のツールの組合せとは、「ユーザが想定していなかったツールの組合せによる分析」または「ツールの作成者が想定していなかったツールの組合せによる分析」を指す。TETDMにおいては、データ型コンバート（可視化モジュールが受け取りを指定するデータ型のデータ以外でも、受け取れるデータ型に自動的に変換してデータの受け渡しを可能にする機能、表1）が実装されており、ツールの組合せの自由度は高くなっている。そのため、あるツールが開発されたときには存在しなかったツールと組み合わせで利用することも考えられ、そのような自由な組合せにおいては、得られる結果を誰も予想していない場合がある。

また複数のパネルを利用する場合、各パネルは基本的に独立になっているため、各パネルに任意のツール

表 1: データ型コンバートのルール

変換前, 後	boolean	int	double	String
boolean	-	0 か 1 に	0.0 か 1.0 に	文字列に
int	0 以外 true	-	(double) に	文字列に
double	0 以外 true	(int) に	-	文字列に
String	"" 以外 true	文字数	文字数	-

をセットして利用できる．そのため、複数のパネルに対するツールの組み合わせ方には限りがなく、むしろ想定外の組合せの方が多くなる．

これらの想定外のツールが組み合わせて利用される場合、意味のない結果が出る可能性もあるが、斬新さや面白さを備えた結果が得られる可能性も秘めており、多様な知見を得る上では一定の効果があると考えられる．

3.3.2 部分テキストの生成

入力したテキスト全体に関する知見を得るだけではなく、特定の条件を満たす一部のテキストに対する知見を獲得したい場合がある．特にデータ量が多くなるに連れ、一部のデータについての特徴を集めることが必要と考えられる．相関ルール分析等の一般的なデータマイニングにおいても、条件部の条件を満たす一部のデータについてのルールを出力した上で、人間が解釈を行っている．

そのため、着目すべき部分テキストを選択する試行錯誤には意味があり、部分テキストの選択肢を提供する支援には意味がある．TETDM においては、形式的には「一部の文あるいはセグメント (段落 or テキスト) を選択」することがテキストの絞り込み (部分テキストの生成) となる．

この一部の文あるいはセグメントを選択する基準として、例えば以下のものが挙げられる．

- 特定の単語 (話題) を含む : ex) 「研究」を含む文
- ある指標の評価値が高い : ex) 句読点が多い文
- 出現位置情報 : ex) 前半の文
- ユーザが指定する文

3.3.3 比較用データの生成

前節で生成した部分テキストに関する特徴を得るためには、単に部分テキストに対する処理結果を得るだけではなく、他の部分テキストに同様の処理を行った場合の結果や、テキスト全体に対する処理結果と比較することは有効と考えられる．すなわち、特徴を理解

表 2: TETDM モジュールのツールタイプ

タイプ名	モジュールの定義
シンプル	データのやりとりが不要で単独で完結する
プリミティブ	一種類のデータ型のデータを 1 つだけ渡す、または受け取る
セミプリミティブ	複数のデータを渡す / 受け取るが、指定モジュール以外との組合せでも動作する
特殊	データに特定の意味があり指定モジュール以外との組合せでは動作しない

したい部分テキスト A に対して、他の条件で絞り込んだ部分テキスト B、あるいはテキスト全体との比較を可能にする．

TETDM においては、複数のパネルを並べることができるため、例えば 3 つのパネルを並べて、それぞれで、部分テキスト A の処理結果、部分テキスト B の処理結果、テキスト全体の処理結果を並べることにしても一定の効果があると考えられる．

しかしその場合、並べられた結果の違いを目で比較する必要があり手間がかかったり、違いを見落とす可能性が生じる．そこでより積極的に比較を促すために、処理結果同士を比較した結果を生成して確認できるようにすることには意味がある．

TETDM では、処理ツールと可視化ツールが分離しているため、各部分または全体テキストに対する処理ツールの処理結果を集めて比較演算を行い、その結果を可視化ツールに渡すことで、既存の可視化ツールに手を加えることなく、比較データを生成して確認することが可能となる．ただし、現時点で利用することができる可視化ツールは、表 2 に示す TETDM の 4 種類のツールタイプのうち、データを 1 つだけ受け渡す「プリミティブ」タイプとしており、前述のデータ型コンバートと組み合わせることで、任意の「プリミティブ」タイプの処理ツールと可視化ツールを、比較用データの生成に用いることが可能となる．

比較演算としては、数値データが得られる場合であれば、減算、max、min などの演算を、文字列データが得られる場合であれば、一致比較 (両方に出現、片方にのみ出現) などの演算を施すことが考えられる．

ここでさまざまな比較方法を用意することで試行錯誤を促し、部分テキストからの知見獲得を助ける．

3.4 結果の解釈

多くのマイニングソフトウェアでは、結果を出力した後は、積極的なユーザ支援を行っていない．高次元のデータを可視化ツール上で理解しやすく表現する方法によって、高次元データを一度に俯瞰してさまざまな知見の獲得につなげることを意図した研究 [小林 13]

はあるが、実際に知見を獲得する過程には踏み込んでいない。一般的な場面では、出力の中から着目する結果を選び、また得られた結果を解釈する際には、ユーザの頭の中、あるいはユーザが用いるテキストエディタなどの上で整理されることが多い。

そこでTETDMにおいては、以下の3つの点についてソフトウェア内で積極的に支援する方法を検討、実装していきたいと考えている。

1. 結果へのアノテーション付け（着眼）
2. 事実への意味付け（解釈）
3. 解釈の一般化（創発）

1. はユーザの着眼支援として、可視化結果の一部に着目を促す仕掛けの導入、2. は結果の解釈支援として、着目した結果に対する意味付けの支援、3. は解釈結果からの創発支援として、集めた解釈を一般化する支援を行う。この結果の解釈の過程は、創造的思考における収束的思考に位置すると考えられる。

3.4.1 結果へのアノテーション付け（着眼）

TETDMにおける可視化ツールの出力結果上に、結果として分かる事実に対して、解釈を与えるべきと考えられる箇所に「」などの印を付けるアノテーション付けを行う機能を設ける。これにより、思考の材料の収集を支援する。

着眼と解釈の手順として、ひとつひとつの着眼と解釈をセットとして繰り返す方法も考えられるが、各着眼点を順次解釈していると、相対的に重要度の低い着眼点に時間がかけられることになったり、またその解釈を行っている間に、他の重要な結果を見落とす可能性も生じる。

そのため、出力結果を全体的に眺めた上で、解釈に値する結果、または解釈を与えるべきと考えられる出力に対して、一通り目印を付け、その後、次節で述べる解釈につなげることが望ましいと考えられる。

TETDMにおいて、具体的なアノテーション付けを実装する方法としては、1) 任意の座標にアノテーション付けを可能にする、2) 可視化ツール内の指定オブジェクトにアノテーション付けを可能にする、という2通りの方法を検討している。

3.4.2 事実への意味付け（解釈）

前節で着眼した結果に対して、その結果に対する意味付けにより解釈を与える支援を行う。着眼点を見ながら、直接その解釈を与えるフォームを用意することによって、直感的に思考できる環境を用意する。すな

わち、着目点が表す「結果」とその「解釈」をペアで入力できるフォームを用意する。

解釈の結果は文字情報として与えられるため、解釈時に分かっていたことでも、時間が経つと意味が分からなく可能性が生じる。そのため、解釈結果をTETDMの出力結果ならびに着眼点と関連づけて記録することで、解釈結果をもとに、その解釈を得るための出力を再現できるようにする。すなわち、各着目点の出力に用いた（部分）テキスト情報と使用したツールの情報は記憶しておき後で再現可能にする。

3.4.3 解釈の一般化（創発）

前節までの手順により得られた、解釈結果を総合的にまとめる支援を行う。複数の解釈結果を端的かつ簡潔に表す支援を行うことで、汎用的な知識の創発を支援する。すなわち、これまでに入力された「結果」と「解釈」のペアを分類、整理して複数の解釈を統合した解釈（知識）を生成する。

TETDM上に、解釈結果の集合を整理できるインタフェースを実装する。複数の「結果」と「解釈」のペアを自由に移動、配置できるようにした上で、関係がある複数のペアを組み合わせることで新たな「結果」と「解釈」を生成してもらうことで、複数の「結果」と「解釈」のペアを統合して、ペアの数を減らす。これを繰り返すことで、最終的に1つの「結果」と「解釈」のペアを作ることを目指す。

4 TETDMを用いた知識創発に向けた予備実験

本章では、前章で述べた知識創発の枠組みに対して、「部分テキストの生成」と「比較用データの生成」について簡易的な実装を行い、行った評価実験について述べる。

部分テキストの生成については、ユーザがフォームから入力した単語を含む段落の集合を部分テキストとする機能を実装した。

比較用データの生成については、実数値配列のデータが生成される場合に限定して実装を行い、実数値列を折れ線グラフとして表示する可視化モジュールを実装した。比較の方法は、二つの実数値列 A, B に対して、 $A - B, B - A, \max\{A, B\}, \min\{A, B\}$ の4つの比較方法のいずれかを選択し、比較用データを生成できる機能を実装した。

12名の大学生、大学院生に対して、26段落(3574字)からなる作家とアナウンサーによる対談テキストを用いて、テキストの分析を行ってもらった実験を行った。テキストを絞り込むための、部分的な話題を表す単語は、

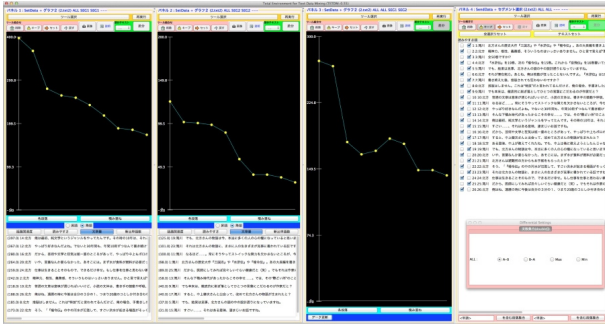


図 5: 知識創発に向けた実験で用いた TETDM の画面

予め 26 段落の 4 割から 7 割程度の段落に出現する単語 10 個を被験者に提示し、使用してもらった。図 5 に、被験者が用いた TETDM の画面を示す。被験者は右端のパネルで、単語を入力することができ、その単語を含む段落からなる部分テキストを 2 種類まで作成することができる。生成する比較データの種類の選択は、右端のパネル上にあるポップアップウィンドウ上で選択できる。左の 3 つのパネルにおいては、各段落の光と影システム [西原 09] による文章の話題関連度、各段落が含む全文字数に対する漢字の割合、段落の文字数、段落で初めて出現する単語数、の 4 つのデータのいずれかを出力する処理モジュールと、与えられたデータを折れ線グラフで表示できる可視化モジュールがセットされている。左の 2 つのパネルは、作成された 2 種類の部分テキストに関するデータを出力し、右から 2 番目のパネルは、左の 2 つの比較データを作成して表示する。

被験者にはまず、文章を読んでもらった後、文章全体についてのデータ（グラフ）をもとに、その文章全体について言える事柄を挙げて、予め用意したエクセルファイルのフォーマットに記入してもらった。その後、指定した個々の話題（単語）についてのデータ、また 2 つの話題の比較データをもとに、話題間の特徴を比較してもらい、再び文章全体について言える事柄を挙げて、エクセルファイルに記入してもらった。

それぞれの回答は上限を 5 個としており、前者で合計 44、後者で合計 48 の事柄が挙げられた。前者の中には、部分的な話題に言及した事柄は 4 つ（2 名）しかなかったのに対して、後者の中には、22（10 名）が部分的な話題に言及するとともに、そのほとんどが複数の話題を比較した内容を挙げていた。このことから、実装した内容が、部分的に扱われている複数の話題に関する新たな事実の列挙、ならびにそれらの解釈を支援する効果があることが確認された。

すなわち、同じ内容を問う課題に対しても、話題によって文章の絞り込みを行う機能（部分テキストの生成）、ならびにその解釈を支援する機能（比較用デー

タの生成）によって、最初とは異なる観点からの考えを得ることができた。そのため本枠組みによって、試行錯誤の幅を広げ、より汎用的な知識の創発につなげられる可能性がある。

5 結論

本稿では、統合環境 TETDM を用いた知識創発支援の枠組みについて述べた。知識創発の枠組みにおいては、試行錯誤と、結果の解釈の 2 つを主な観点として挙げ、そのそれぞれに対する実現の方法について述べた。試行錯誤における部分テキストの生成と比較用データの生成に関して簡易的な実装と評価実験を行い、試行錯誤の選択肢を広げることが、幅広い解釈につなげられる可能性を示唆した。

本稿の実験では、解答をエクセルファイルに記入してもらった形で行ったが、今後は可視化結果への着眼から知識創発に至る論理的解釈を支援する枠組みを実装し、その上で評価を行っていきたいと考えている。

参考文献

- [Guilford 59] Guilford, J. P.: Three faces of intellect, American Psychologist, Vol.14, No.8, pp.469 – 479 (1959)
- [小林 13] 小林弘明, 三末和男, 田中二郎: 色による特徴表現を用いた高次元データの可視化, 情報処理学会研究会報告, Vol.2013-HCI-152, No.23, pp.1 – 8 (2013).
- [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).
- [砂山 11] 砂山渡, 高間康史, Danushka Bollegala, 西原陽子, 徳永秀和, 串間宗夫, 松下光範: Total Environment for Text Data Mining, 人工知能学会論文誌, Vol.26, No.4, pp.483 – 493 (2011).
- [砂山 13] 砂山渡, 高間康史, 西原陽子, 徳永秀和, 串間宗夫, 阿部秀尚, 梶並知記: テキストデータマイニングのための統合環境 TETDM の開発, 人工知能学会論文誌, Vol.28, No.1, pp.1–12 (2013).
- [砂山 14] 砂山渡, 高間康史, 西原陽子, 梶並知記, 串間宗夫, 徳永秀和: 統合環境 TETDM を用いたマイニングツールの開発と利用の実践, 人工知能学会論文誌, Vol.29, No.1, pp.100–112 (2014).