博士論文

時間の経過に着目した
蓄積文書集合処理の研究

田中 克明

目次

図目次

表目次

第 1 章

序論

1.1　背景と目的

本研究の目的は、文書集合の理解に時間経過の視点を導入することである。

1.1.1　文書集合と情報検索・情報分類

人間は、活動に伴い、数多くの文書を作成する。ある文書は活動を記録するために作成され、ある文書はその作成自体が活動の目的である。これらの文書は時間の経過に伴って数が増えていき、蓄積され、文書の集合となる。

例えば、日本における公文書管理制度に基づく国の行政文書ファイル等の数 [64] は、2011 年度には約 1,500 万件、新規作成分を順次加算すると 2019 年度には 3,700 万件である（図 1.1）。あるいは、Wikipedia 日本語版の記事数 [2] は、2003 年 12 月には約 2 万件、2020 年 12 月には約 124 万件である（図 1.2）。

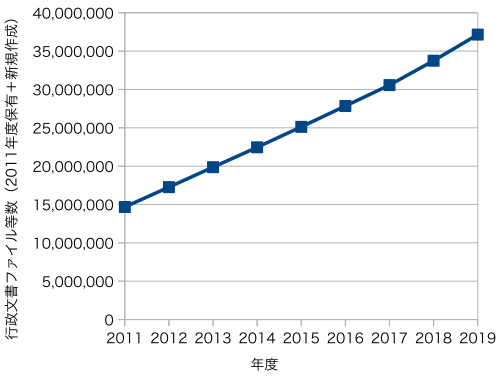


図 1.1 行政文書ファイル等の数（[64] より作成）

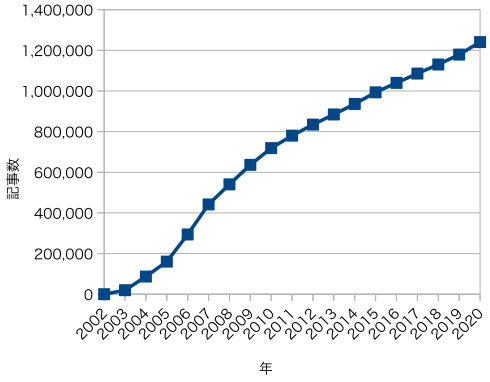


図 1.2 日本語版 Wikipedia の記事数（[2] より作成）

文書集合の内容を理解するために、情報検索や情報抽出・分類の研究がなされている。情報検索は、Google に代表される検索エンジンのように、キーワードなどに基づき利用者の意図との一致度合いが高い文書を探し出す手法である。情報抽出・分類はトピック抽出や文書クラスタリング [8][25] のように、文書集合中に記述された特徴的な内容の抽出や類似する文書・内容のグループ化を行い、文書集合のおおよその特徴を把握する手法である。

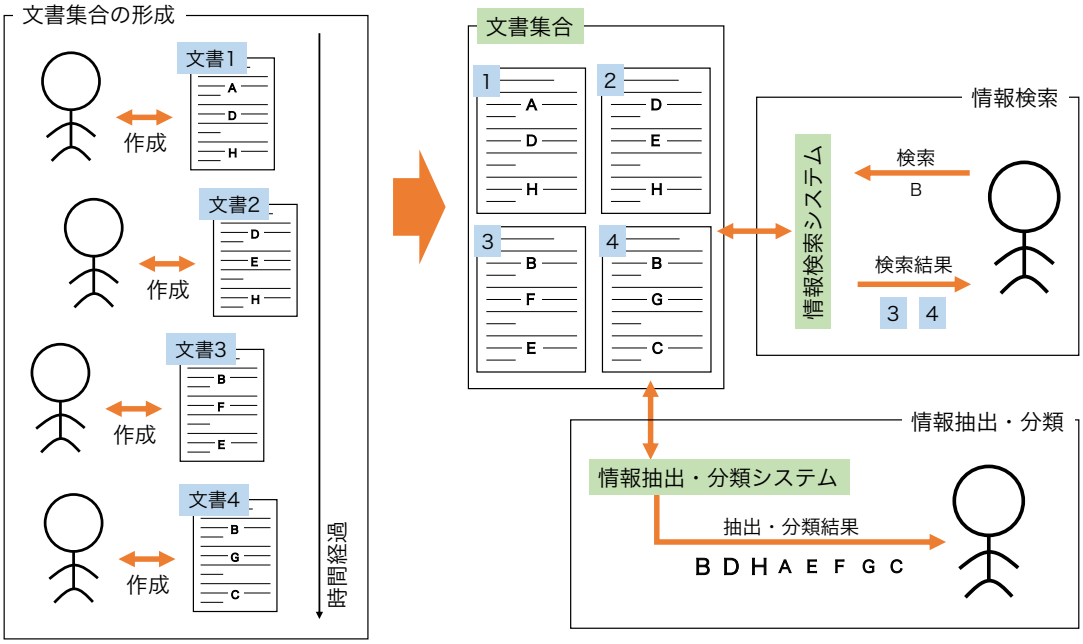


図 1.3 文書集合と情報検索／情報・抽出分類

図 1.3 に文書集合と情報検索と情報抽出・分類の関係を示す。文書集合が作成された後に、文書集合を利用する人間が検索システムや情報抽出・分類システムを操作し、文書集合の内容の把握を行う。

図 1.3 の左側に示したように、文書が時間の経過に沿って作成・蓄積され、文書集合となる。文書集合中には、独立して作成された新規の文書のみでなく、以前の文書の記述内容を参照する文書や以前の文書の内容を更新する文書が含まれ得る。しかし、情報検索や情報抽出・分類では、それらを適用する時点までに蓄積された文書の集合、あるいは、過去のいずれかの時点で得られた文書の集合を対象する。文書の蓄積過程には注意を払わず、その結果、文書集合に含まれる記述内容が、どのような経緯によりその時点の形に至ったのかは、各手法の結果に反映されない。

1.1.2 文書集合と時間経過に伴う変化

文書集合は、時間の経過に沿って作成され、作成された文書が蓄積されることにより形成される。1.1.1 節にあげた行政文書や Wikipedia の記事も、図 1.1、図 1.2、に示したように、時間の経過に沿って文書が作成・蓄積されることにより、文書集合が形成される。

時間の経過に沿って文書集合と文書を扱うと、どのようなことがわかるだろうか。例として、人工衛星の設計過程を観察して記録した文書の集合を考える。この文書集合中に、

(A) 太陽電池の試験を行った
(B) 無線機を購入した
(C) 無線機の試験を行った
(D) 無線機の機種を決定した

という内容の文書が存在すると考えてみる。

情報抽出・分類の手法を使うと、例えば、「太陽電池」「無線機」のようにどのような種類の機器が存在するか、あるいは、「無線機」に対してどのような状態（「機種の決定」「購入」「試験」）が存在するか、を読み取ることができる。このような情報は、新たに機材を購入し、利用する際に、「どのような機材が必要か」の決定を行う際に役に立つ。

一方、文書が作成された時刻に着目すると、例えば、(B) と (C)、(D) の関係から、「購入から試験の実施、機種決定へ」という作業の過程が読み取れる。このような時間の経過に沿った情報は、設計を進める際に、「どのように作業を進めるか」の決定を行う際に役立つと考えられる。

本研究では、これらの後者、時間の経過に沿った観察により得られる情報に着目する。

1.1.3 本研究の目的

これまでに述べたように、情報技術の発達に伴い、さまざまな場面で文書が作成され、記録、蓄積されている。文書集合に含まれる内容は、情報検索の手法を用いてそれらを検索し、必要な文書を参照することが可能である。また、文書集合を情報抽出・分類の手法を用いてそれらを分類し、俯瞰することも可能である。

しかし、これらの手法では、文書集合中の文書間にまたがる時間の経過を考慮しない。その結果、記述内容の移り変わりにも注意を払わない。これは、作業や作業対象の変化の過程は気にせずに、作業対象の種類や状態を細かく分解して見ていることに相当する。

時間の経過を考慮し、文書集合に含まれる記述内容を時間の経過に沿って整理すれば、記述内容の移り変わりを確認することが可能である。先述した設計過程の例でいえば、「ある部分にどのような作業を行ったか」という視点を得ることができる。

そこで本論文では、文書集合の形成過程における時間の経過に着目し、文書集合に含まれる記述内容の時間経過に沿った変化の過程を、文書集合の利用者にあわせて解きほぐしながら提示するシステムを提案し、文書集合を時間の経過に着目して扱うために必要な手法について、研究を行う。

本研究の前提と目的は、以下の通りである。

前提として、本研究では、文書集合を利用する個人が存在し、この個人が、文書集合の俯瞰的な把握にとどまらず、興味を持った個別の内容を理解し、何らかの情報を得ようとする状況を考える。

文書集合の内容を理解しようとする個人がいることから、扱う文書集合は「Web 全体」のように漠然としたものではなく、「ある機器の設計過程の記録」のように、何らかの形で集合としての境界が存在するものとする。また、文書集合の内容が容易に理解できるようであれば、とりたてて研究の対象とする必要もないため、蓄積期間が数か月以上の比較的長期間にわたるなど、ひとりの人間ではすべてを読みつくすことが難しく感じられる量の文書からなる文書集合を研究対象とする。例えば、目的を共有するグループの作業記録、継続的に開催される会議の議事録、同一のトピックに関する SNS 上の書き込みなどを、本研究では扱う。

目的の 1 つめは、時間の経過と文書集合の関係を整理し、文書集合内における時間経過に沿った変化を抽出する手法の確立である。そのために、文書集合とそこに含まれる時間経過の関係について整理し、第 2 の目的として述べるシステム構築の指針とする。

目的の 2 つめは、時間の経過に沿って文書集合を扱う枠組みの構築である。すなわち、文書の集合における時間の経過に着目し、文書集合に含まれる時間経過に沿った変化の過程を、文書集合を扱う人間の着眼点に合わせて提示するシステムを構築する。

本論文では、第 1 の目的の時間の経過と文書の関係の整理結果によって、第 2 の目的に述べたシステムを構築し、このシステムの適用結果により、あらためて時間の経過と文書の関係を整理する。

1.2 本研究の概要

1.2.1 文書集合からの時間経過に沿ったトピック抽出と操作システム構築

本研究では、時間の経過に沿って文書集合を扱うことを目指す。そのために、蓄積された文書を時間の経過に沿って整理し、それらに含まれる変化の過程を扱う仕組みを導入することにより、人間による文書集合における内容変化の把握を支援するシステムを構築する。具体的には、蓄積期間が数か月以上の比較的長期間にわたり、ひとりの人間ではすべてを読みつくすことが難しく感じられる量の文書集合に対し、時間の経過に沿って大まかな構造を与え、読み手が着目した点に応じて構造を再構成し提示するシステムを構築する。概要を図 1.4 に示す。

このシステムではまず、文書集合から、トピック抽出の手法を応用し、時間経過に沿ったトピックの遷移を抽出する。得られたトピックの遷移を、トピック遷移構造と呼ぶことにする。続いて、情報検索と同様に、システムの利用者が文書集合中から把握したい内容を示す単語を指定し、トピック遷移構造のうち、指定単語に関連がある部分構造を結果として返す。これにより、

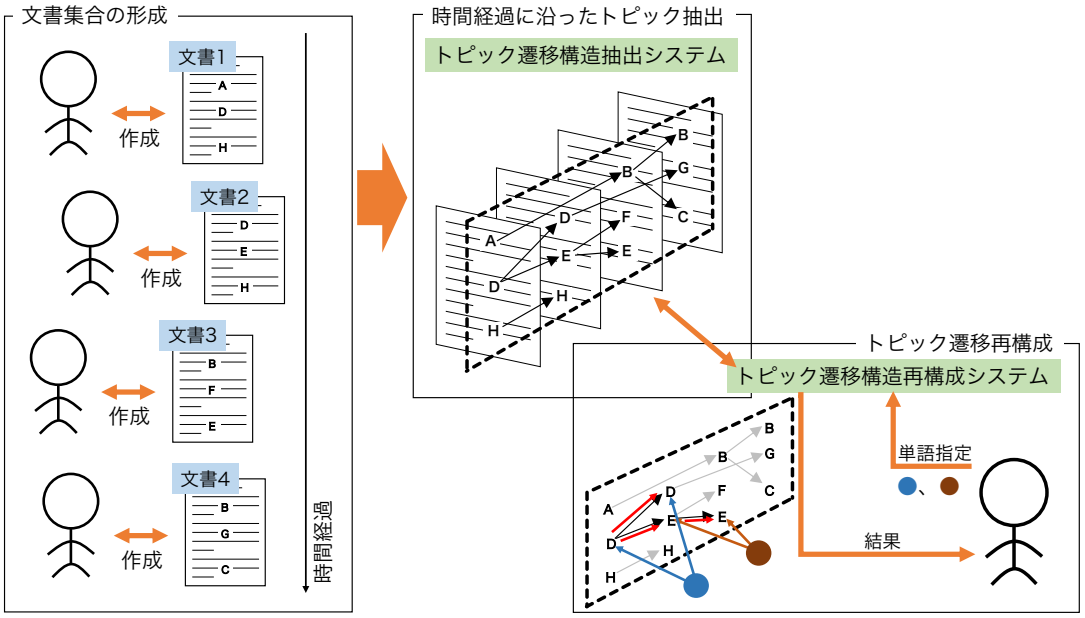


図 1.4 時間経過に沿ったトピック抽出と再構成

システムの利用者が、着目した事柄が文書集合中でどのような経過を経ているかの把握を行うことが出来る。

システムの具体的な構成や処理手法を 4 章に、システムの適用事例を 5 章、6 章に述べる。

1.2.2　文書集合における時間経過の検討

本研究では、時間の経過に沿って蓄積された文書の集合を扱う。しかし、「時間の経過に沿って蓄積された文書集合」にも、さまざまな形態が考えられる。本論文も、各部分部分は別の時刻に筆者が記したものであり、時間経過に沿って蓄積された集合と考えることができる。SNS の書き込みを集めた集合も、さまざまな人がさまざまことを記した文書の集合である。

本論文であれば書き手は筆者 1 人であり、SNS の書き込みであれば書き手は多様な人々である。書かれた対象も、本論文であれば文書集合と時間経過に関する考察だが、SNS であれば多様な事象である。また、本論文であれば、新たに書き加える際にはそれまでに書いた内容を踏まえる。一方、SNS に文書が記される際には、以前に書いた内容を踏まえる場合、社会的背景を踏まえる場合、個人の心情のみを踏まえる場合など、さまざまな場合がある。

このように、文書集合の蓄積のされ方により、文書集合は異なる性質を持つと考えられ、本研究ではこれらの違いについて、まず、3 章にて検討を行う。前後するが、1.2.1 節にて述べたシステムは、この検討に基づき構築を行う。さらに、システムのを構築し、システムの利用結果を踏まえ、6 章でも、あらためて同様の検討を行う。

さらに、本研究では、前述したとおり、時間の経過に沿った観察により得られる情報、すなわ

ち時間の経過に沿って起きた変化に着目する。「何かが変化したということは、なにか変化しないものがあることが必須である」[74] とあるように、時間の経過に伴う変化を観察するためには、実時間が経過すること以外に、文書集合に含まれる変化の中にあって「なにか変化しないもの」が必要である。そこで、文書集合から、時間の経過に沿って起きた変化の抽出を支援するために、「なにか変化しないもの」を「通時的対象」として抽出する手法について、7 章にて述べる。

1.3 本論文の構成

第 2 章では、時間の経過に沿って情報を扱う先行研究を取り上げる。時間の経過に沿った情報の可視化を行う研究、時間に沿って蓄積された文書集合からの情報分類・情報抽出の研究、時系列データ解析、ベイジアンネットワークを用いた因果モデリング、設計支援システム、深層学習による系列処理を取り上げる。また、対象とする「時間の経過」の長さなどにより先行研究を分類し、本研究の位置づけを行う。

第 3 章では、文書集合と時間経過の関係、文書集合の記述対象の関係について、整理と検討を行う。文書は何らかの対象について人間が記述を行ったものである。複数の文書からなる文書集合の記述対象は、時間経過の中で独立して存在する場合、時間経過に伴い変化する場合などが考えられる。また、ある事物を記述対象とする場合、事物への人間の働きかけを記述対象とする場合など、記述対象の性質も複数考えられる。これらを蓄積のされ方の形態などを踏まえて整理し、その特徴を検討する。

第 4 章では、文書集合から時系列に着目してトピックの移り変わりの抽出を行い、抽出結果をインタラクティブに操作し、文書集合の理解支援を行うシステムについて述べる。文書の作成された実時間により文書集合を複数の部分集合に分け、古い内容が忘却される過程を取り入れつつ、時間経過に沿ったトピックの抽出を行う。続いて、抽出したトピックを俯瞰表示するシステム、単語の出現率の推移を表示するシステム、指定した単語を含む部分集合を表示するシステム、元の文書を表示するシステムなどについて述べ、それぞれの動作を確認する。

第 5 章では、第 4 章にて述べたシステムの実際の文書集合への適用事例と、その結果を述べる。文書集合として、大学における小型人工衛星の設計プロジェクトの議事録を取り上げる。設計過程で検討された事項、例えば、使用する部品の変遷や、ある部分への作業内容の推移が抽出できることを述べる。

第 6 章では、第 4 章にて述べたシステムを、性質の異なるいくつかの文書集合へ適用し、文書集合と時間の経過についてあらためて検討を行う。文書集合として、政府の審議会の議事録、Twitter 上から集めた同一のキーワードを含むツイートの集合を取り上げる。検討の結果、文書集合の内容把握は可能だが、第 5 章と比較すると、時間の経過を確認しづらいことが確認された。この理由に関する考察も述べる。

第 7 章では、文書集合から時間経過に沿った変化を抽出する手がかりとして、「通時的対象」の概念を定義し、通時的対象を文書集合から抽出する手法を述べる。通時的対象は、1.1.2 節にて述べた「無線機」のように、文書集合中で時間経過を通して変化していく対象である。そこで、共起する単語が「購入」「機種」のように類似するが変化する単語を、通時的対象の候補として抽出し、頻出語と比べ通時的対象である割合が高いことを確認した。

第 8 章では、第 4 章から第 7 章を通して文書集合を時間の経過に着目して処理した結果をまとめ、工学的に利用価値のある情報の獲得について、従来研究との比較をしつつ検討を行う。また、時間の経過が有効に働くように文書集合を蓄積するためには、どのような形態での文書の作成・蓄積が望ましいか、記述の対象、記述を行う人間、記述される文書の関係を整理する。

第 9 章では、本論文のここまでの議論を踏まえ、論文全体のまとめと今後の展望を述べる。

第 2 章

関連研究と本研究の位置づけ

本章ではまず、本研究に関連した研究として、時間の経過に着目して情報を取り扱う先行研究について述べる。続いて、各研究が扱う時間の長さと、各研究の処理結果を人間が理解する際に処理前のデータへの理解をどの程度必要とするかの 2 つの観点から、本研究との関連をまとめる。

まず、以下の順に、時間の経過に着目して情報を扱う研究を取り上げる。

1. 時系列に沿った情報の可視化
2. 時系列トピック抽出
3. ベイジアンネットワークと因果関係モデリング
4. オントロジー
5. 設計支援システム
6. 時系列解析
7. その他の時間の経過に着目した研究

2.1　時系列に沿った情報の可視化

代表的な Web 検索エンジンである Google は、検索結果を時系列にあわせて表示するサービス Timeline View を 2007 年から 2012 年まで提供していた（2020 年時点では提供されていない)。Timeline View は、ユーザのキーワード入力に対する検索結果を、年代順の一覧と年代別の出現数のグラフとして提示する機能であった。例を図 2.1 に示す。

なお、現時点での Google は、Web ではなく書籍を対象として単語の出現比率の時間経過に沿った変化を提示する Google Books Ngram Viewer[20] や、検索対象ではなく、検索のために入力された検索キーワードの時系列変化を示す Google トレンド [21] を提供している。

時間経過に沿った情報の可視化には様々な手法が提案されており [5]、人と人の関係の変化の

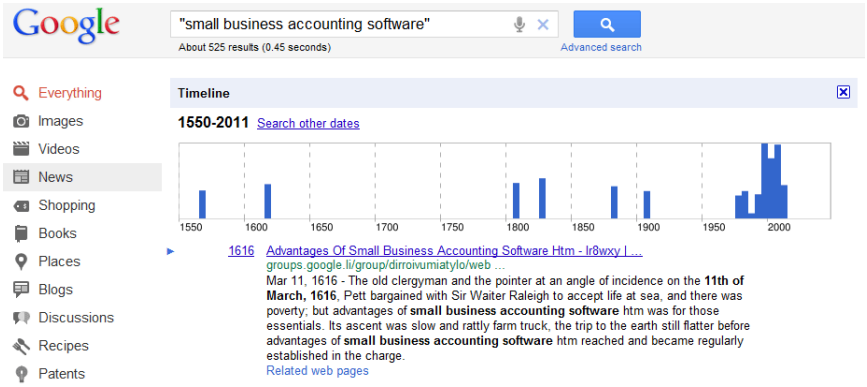


図 2.1 Google Timeline View[36]

可視化手法に関する研究 [16]、物語内の空間移動の可視化手法の研究 [32] などが行われており、2 次元、2 次元平面を複数並べた 2.5 次元（図 2.2）などでの可視化手法が検討されている。

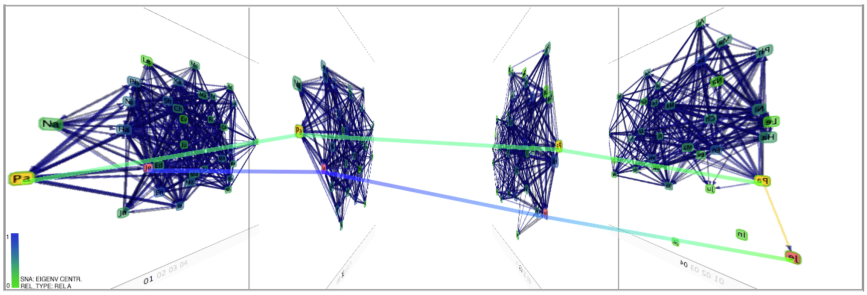


図 2.2 ネットワークの時間に沿った変化の 2.5 次元表示 [16]

2.2　時系列に沿ったトピック抽出

2.2.1　Dyamic Topic Models

時系列に沿ったトピックの抽出手法として Dynamic Topic Models[7] や、これを発展させた Dynamic Embedded Topic Model[14] などがある。Dynamic Topic Models は、LDA などに代表されるグラフィカルモデルを、時系列情報に適用し、時間経過を含む情報に含まれる関係性の抽出を試みる研究のひとつである。

Dynamic Topic Models による時系列に沿ったトピック抽出の例を図 2.3 に示す。図 2.3 は、長期間にわたる論文から抽出したトピックの推移を俯瞰的に示したものである。

Dynamic Topic Models では、トピックは対象とする文書集合が生成された全期間にわたって存在するものとして、存在を固定して扱い、それらのトピックの出現確率がどのタイミングで大きく、どのタイミングで小さいかの計算を行う。なお、トピックの出現確率を 0 とすることで、存在しないものとして扱うことも可能である。

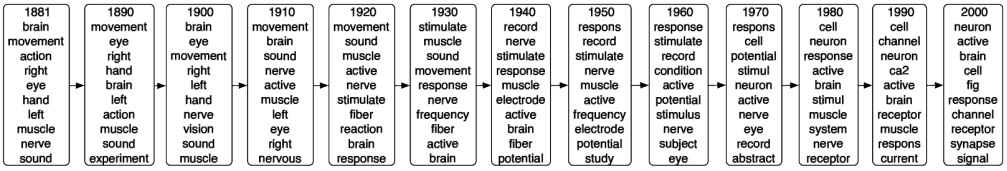


図 2.3 Dynamic Topic Models による抽出例 [7]

2.2.2 ニュース記事からの話題の抽出と追跡

Topic Detection and Tracking (TDT)[6] など、新聞記事や放送の書き起こし文字列から、「イベント」に関連する部分を抽出することを目的とした研究が行われている。TDT では、以下の 3 つのタスクを設定し、それぞれに対し学習データと正解データ付きのデータを用意し、評価を行っている。

1. Tracking Task：既知のイベントに言及している文書を逐次抽出する
2. Detection Task：新しいイベントの発生を抽出する
3. Segmentation Task：放送記事など一続きの文書を内容ごとに区切りをつけ分割する

TDT の特徴は、対象がニュース記事であることと、出来事を示す「イベント」に対応する文書の特定と同じイベントに関連した記事の追跡を行うことにある。同様に、ニュース記事や SNS などの文書集合から出来事（イベント）に関する記述を時系列に追って抽出する研究はいくつかなされている [44]。また、イベントの抽出・追跡などの成果をどのように応用するかは TDT の範囲には含まれないが、例えば、各タスクの成果を応用して、記事の一覧を可視化するシステムがいくつか提案されている [17]。

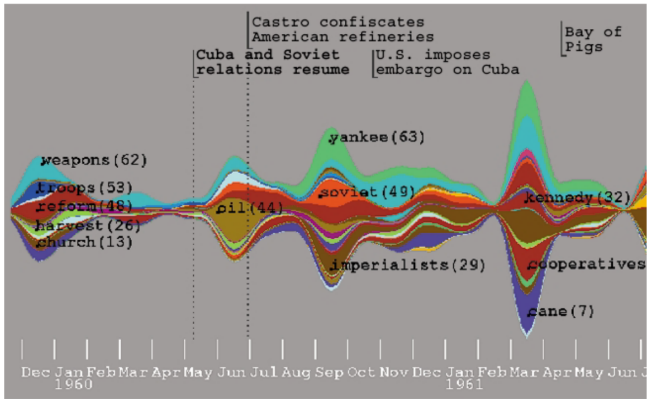


図 2.4 ThemeRiver の例 [22]

その他、ニュース記事からのトレンド抽出を目的とし文書クラスタリングを行う手法として採

用し、古い文書を作成からの経過時間にそって忘却するモデル [50] も提案されている。ニュース記事の内容の時間経過を扱う研究では、特定の内容の抽出ではなく、ニュース内容の可視化を行うことを目的とした研究 [22]（図 2.4）も行われている。

2.3　ベイジアンネットワークと因果関係モデリング

ベイジアンネットワークは、ある概念や事象をノードとし、それらの間の遷移確率を示したネットワークである。この中で、時間に着目しつつ因果関係についてベイジアンネットワークを構築し、これを利用することでサービス可能知識とする研究が行われている [71][72]。

ベイジアンネットワークでは、ノードとなる「事象」をネットワーク構築者が設定し、ノード間の遷移確率を計算機により求める。すなわち、どのような因果関係をモデル化するか人間が選択を行い、観測された事象により選択された事項間の遷移確率を求める。

ある事象に対する別の事象の発生確率の推測を行うためには、その事象のノードから次のノードへの推移確率にそって、ネットワークをたどっていけばよい。このように、ベイジアンネットワークには状況依存性を持たせることができ、十分な情報に基づいてベイジアンネットワークを構築すれば、局面に応じた状態遷移の提案を得ることが可能となる。例えば、ベイジアンネットワークをサービス可能知識として用いるオープンライフマトリクスプロジェクト [72] では、子どもが遊んでいる環境から、子どもの行動と遊んでいる環境要因などに着目して、それらの要因と発生してしまう事故との間の因果関係をベイジアンネットワークを用いて図 2.5 のようにモデル化し、事故によるリスクを下げる研究などが行われた。

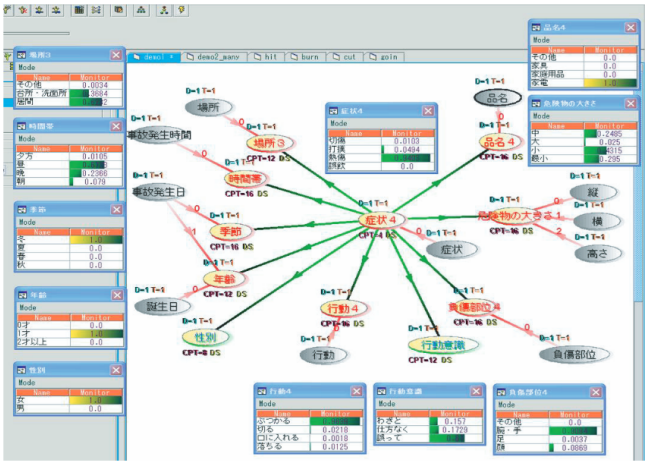


図 2.5 子どもが遊ぶ環境の事故事例から構築したベイジアンネットワーク [71]

2.4 オントロジー

オントロジーは、「問題解決システムを構築する際に用いられる概念の体系的記述」[73] であり、問題解決に伴う概念を体系的に整理しておくことにより、さまざまな概念と利用されるべき状況との対応付けを可能とすること、すなわち、体系を構築しそれに沿って知識をあらかじめ整理しておくことを目的とする。オントロジーに関する研究は、構築手法に関するものから、構築と利用をあわせて包含した知識集約の方法論まで、広範囲に渡る。

オントロジーを用いつつある対象について時間変化を含んだモデルを扱う研究として、人工物の機能を中心とする技術知識を統合的に管理し活用する環境 OntoGear[57] がある。OntoGear では、対象とする人工物について、その機能とそれを実現するための構造を示す機能分解木、および時間軸として機能発揮時間軸と装置変化時間軸を提案している。OntoGear 中で、対象物の機能・構造を表す機能分解木は、装置変化時間軸にそって変化するものとして扱われ、時間軸上に並べた機能分解木を俯瞰したり、対象物を扱う人間が、どの時点のどの機能・構造について議論を行っているかを示したりすることができる。

また、時間に関するオントロジーとして、原因と結果の因果関係と時区間概念を結びつけたオントロジー [28] の研究がなされている。

2.5 設計支援システム

設計は、人間が作りたいものとの間で時間をかけてインタラクションを行う過程であり、時間の経過に沿って行われる。

設計は、しばしば設計の対象がどれだけ具体的になったかに着目して段階わけが行われる。例えば機械設計ではおおよその構造を検討する概念設計、構造関係を決定する基本設計、対象の図面を記す詳細設計と、順に進んでいくように分けることができ、これらの段階に応じた設計支援システムが存在する。

一方、工程によらない分類も可能である。たとえば、1.2.2 でも述べたように、扱う対象により、設計対象となる構造を扱う情報をプロダクト情報、設計者による設計活動に関する情報をプロセス情報と分類できる [53]。

プロダクト情報の表現方法には、例えば日本工業規格における機械製図（JIS B 0001）、デジタル製品技術文書情報（JIS B 0060）のように記述方法の規格が存在する一方、プロセス情報の表現方法、つまり設計プロセスにおいて何をどのように表現し記録し活用するかについては、規格の制定までは行われていない。

2.5.1　プロダクト情報を扱う設計支援システム

コンピュータを用いた設計支援システムとして、Computer Aided Design（CAD）システムが広く用いられている。CAD は、物の製造に必要な図面を計算機上で作成、管理することを主な目的としたシステムである。また、前述したプロダクト情報を扱う典型的なシステムでもある。

現在の CAD は、図面という 2 次元あるいは 3 次元上の座標とその関係の集合を扱うだけではなく、図面のある場所をどのような目的のためにその形状にしたのか、という情報を付加して取り扱うことができるようにする拡張も行われている。たとえば CATIA[46] では、ナレッジ・テンプレートとして、ある部品の設計意図を形状情報と同時に記録しておくことにより、同一の部品を同一の意図で別の箇所で用いる場合に、部品の形状を調整する機能を持つ [47]。このためには、設計者が、設計の対象物を、周囲とのどのような関係性によりその形状にしたか、というように、設計の意図するところを注意深く書き下す必要がある。

このように、ある部分が周囲とどのような関係を持つかなど、設計物の構造についての定式化を行えれば、現実世界に対象が存在しなくても、対象のふるまいを計算機上でシミュレーションが可能である。

また、設計対象の設計結果が設計解として満たすべき要求の定式化ができる場合、計算機による評価を組み合わせ、設計過程において行う必要があるさまざまな決断を行った結果に応じた、それぞれの場合の設計解候補を探索させることができる。さらに、シミュレーションを利用して、設計解候補の関係を可視化し設計者に提示することにより、設計上の問題そのものの特性、たとえば、明らかになっていない要素間の因果関係などの情報を設計者に提供する手法も研究されている [59]。

2.5.2　プロセス情報を扱う設計支援システム

次に、人間による設計活動を対象とした支援、主にプロセス情報を扱う支援について述べる。このような支援としては、例えば、設計がなされた際に根拠となる情報を知識として得られるよう、設計根拠の記録と活用に関する研究がなされている。設計根拠を獲得するためのシステムとして、議論の記録を行う gIBIS[11]、Compendium[37] などが提案されている。例えば gIBIS は、設計初期の議論において、議論内容を Issue、Position、Argument に分類し、それらの関係を画面上で整理しながら設計と議論をすすめるシステムである。

他にも、設計活動を扱うためには、必然的にその対象である設計対象も扱うことから、設計対象のモノと人間の設計活動を統合的に扱い、知識や情報が生成されていく過程に着目した時系列の視点と、問題と代替案の関係に着目した論理的な視点のそれぞれに立脚して、設計支援を行う研究 [61] も行われている。2.4 節にて述べた OntoGear も、時間軸として設定している機能発揮

時間軸は設計対象を扱い、装置変化時間軸は設計活動を扱うものととらえることができ、設計活動と設計対象を統合的に扱った研究である。

2.6 時系列解析

時系列解析とは、時間の経過に沿って連続して観測可能な値、例えば気温や株価、機器の監視データなどに対して、時系列モデルを想定し、そのデータの変化の予測、異常検知、類似パターンの発見などを行う。時系列解析では主に、時間の経過によりデータの確率分布が変化しない定常的な過程を対象とする。

モデルを利用する手順は、時系列モデルの特定化、モデルの推定、モデルの検証・診断、モデルを用いた予測といった流れに整理される [9][63]。時系列モデルは、時間差の増加に伴う系列相関の減衰が速い短期記憶モデルと、時間的な従属性が強くデータ間の時間差に伴う系列相関の減衰が少ない長期記憶モデルに分類することができる。

また、時系列データを表層的にモデル化するにとどまらず、データを生成するシステムの振る舞いを推定する動的システム学習 [70] の研究も進められている。

2.7 その他の時間の経過に着目した研究

2.7.1 プロセスマイニング

情報システムなどのイベントログから、作業の手順などの発見を目指すのがプロセスマイニング [65][41][42] である。イベントログはある事例のアクティビティに伴って発生したイベントの記録からなる順序集合であり、イベントごとにアクティビティ、事例との関連や発生時刻などの情報を記録したものである。離散的なイベントログをもとに、プロセスを実行する組織内の役割（ロール）やロール間の関係、アクテビティの実行に要する時間、プロセス内の条件分岐のルールなどの知識を抽出し、実際のプロセスを発見、監視、改善することを目指す。

2.7.2 深層学習による時系列処理

深層学習において時系列データを扱う代表的な仕組みが、Reccurent Neural Network（RNN）である。例えば、RNN により自然言語で記述された文を扱う場合、単語の情報を入力として与え、1 単語ごとに RNN を動作させ、中間層の出力を次の単語の入力に追加する。これにより、ある単語より前の入力が、それ以後の入力に影響を及ぼすことができる。

RNN の中間層から中間層への直接伝播では、長期間に渡った情報の伝達が難しい。そこで、Long Short Term Memory（LSTM）[24] と呼ばれる記憶領域を用いる方式から、Attention[30]

と呼ばれる、伝播させたい情報を別に扱う仕組みが提案され、Attention の伝播を文の前から後ろ・後ろから前の双方向へ拡張した BERT[12] などの単語列の予測機能を持つ言語モデルが、自然言語処理の課題解決に広く用いられている。

ニューラルネットワークによる自然言語処理では、単語のならびである文を対象とし、単語をネットワークへの入力とする。

2.7.3　コンセプトドリフト

機械学習において、扱うデータの置かれた文脈が変化することにより分類対象の概念にも変化が生じる事象を、コンセプトドリフトという [45]。例として、天候の予測方法の季節による違い、購買者パターンの曜日や代替品の存在、インフレ率などによる変化があげられる [40]。コンセプトドリフトに対応する方法として、学習データを時間経過に伴い忘却させる、変化を検出する、学習モデルを差し替えるなどの手法が研究されている [19]。

2.8　本研究と従来研究の関連性

本節では、ここまでに述べた従来研究と、本研究の関連について、研究対象とする時間の長さと、情報の扱い方の違いについて述べる。

2.8.1　対象とする時間の長さ

2.2 節に述べた時系列文書集合からのトピック抽出では、複数の文書に渡るトピックの関係を扱う。一方、2.7.2 節で述べた深層学習による時系列処理では、文章内の連接する単語間の関係を中心に取り扱う。前者が扱う時間の長さは、ニュース記事間の関係を扱う場合を考えると、数日から長ければ数年である。一方、後者が扱う時間の長さは、人間が読む速さや発話する速さからに合わせて考えると、数秒から数分程度である。このように、時間に着目して情報を扱うシステムでも、扱う時間の長さには大きな違いがある。

本章で挙げた先行研究は、おおまかに以下のように分類することができる。

1. 長い時間経過を扱う研究
   時系列に沿ったトピック抽出、ベイジアンネットワークと因果モデリング、オントロジー、設計支援システム、コンセプトドリフト、時系列解析（長期記憶）
2. 短い時間経過を扱う研究
   プロセスマイニング、深層学習による時系列処理、時系列解析（短期記憶）

本研究では、2.2 節の時系列文書からのトピック抽出同様、複数の文書の間にまたがる記述内

容の遷移を対象とし、比較的長い時間の経過に着目して情報を扱う。3.1 節で、あらためて研究対象とする時間の長さについて検討を行う。

2.8.2 処理データに対する事前知識の必要性

2.2 節に述べた時系列に沿ったトピック抽出手法では、トピック抽出の元となる文書集合として、多くの人が共通してあらかじめ知っているであろうニュース記事や論文を対象としている。この理由として、抽出の結果として得られた図 2.3 や図 2.4 のような俯瞰的な表示を人間が理解するためには、理解する側の人間が、文書集合に含まれる内容について事前にある程度知っている必要があることが挙げられる。

一方、2.3 節の因果関係モデリングでは、時系列データをベイジアンネットワークとして表現する際に、ノードとなる部分、すなわち因果を構成する出来事を人間があらかじめ指定することによりデータから学習を行い、人間が再利用可能な知識とする。すなわち、データの構造化に人間が関与して整理を行うため、構造化結果を利用する人間は、データに関する事前知識を持っている必要はない。例えば図 2.5 のベイジアンネットワークの利用者は、子どもの遊ぶ環境における事故例についての事前知識を持っている必要はない。

このように、時間経過に着目して情報を扱う研究でも、あるデータを処理して何らかの構造を作る手法で得られた結果に対し、これを理解する人間が、処理前のデータに関する知識をどの程度必要とするかに違いがある。この違いは、処理結果を理解するためのインタラクティブな仕組みが各研究の枠組みの中に含まれているか否かにより起きると考えられる。

本章で挙げた先行研究は、おおまかに以下のように分類することができる。

1. 利用者が処理前のデータについて知っている必要がある研究
   時系列に沿ったトピック抽出、設計支援システム、時系列解析、プロセスマイニング、コンセプトドリフト
2. 利用者が処理前のデータについて詳しく知らなくても良い研究
   ベイジアンネットワークと因果モデリング、オントロジー、深層学習による時系列処理

本研究にて構築するシステムでは、システム利用者が処理対象とする文書集合に関する知識を持つことを前提としない。このために、文書集合を処理した後、俯瞰的な表示を行うにとどまらず、利用者が処理結果をインタラクティブに操作するための仕組みを設ける。

2.9 本章のまとめ

本章では、本研究同様に時間の経過に着目して情報を扱う先行研究について述べた。また、先行研究について、数日から数年単位の長い時間経過を扱う研究、数秒から数分の短い時間経過を

扱う研究、および、各研究による手法の処理結果を利用する人間が、処理前のデータについて知っている必要性がある研究、詳しく知らなくても良い研究に分類を行った。

これに基づき、本研究では、長い時間経過を扱い、構築するシステムの利用者に処理前の文書集合に関する知識を求めないことを述べた。

第 3 章

時間経過と文書の蓄積のされ方

本論文では、文書間の時間の経過に着目し、文書集合に含まれる時間経過に沿った変化の過程を提示するシステムを提案し、文書集合を時間の経過に着目して扱うことにより得られる情報の研究を行う。

本章では、文書集合の処理に先立ち、文書集合の蓄積のされ方を確認し、文書に記された対象と文書を記した人間の関係、文書の蓄積過程と時間経過の関係などについて、検討を行う。

3.1　時間の経過と文書の作成・蓄積過程の関係

本節では、時間の経過と文書の作成・蓄積の過程について検討を行い、本研究が、文書集合内におけるどのような時間の経過に着目するか、明らかにする。

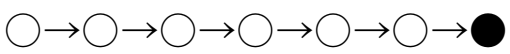
3.1.1　作曲と演奏

フッサールは、人間の持つ時間意識を形成しているのは記憶であると述べている [48]。フッサールによると、人間は音楽を聴いているとき、聴こえた音を順次記憶し蓄積することで、ひとつずつ聞こえてくる音の連なりとして「過去把持」を形成し、これによりバラバラの音を曲として認識する。すなわち、音を記憶することで、時間の経過を認識している。また、音の連続から曲の認識に至る過去把持を第一次記憶と呼び、これに対して、音楽を聴き終え、過去に聴いた音楽を想起することを第二次記憶と呼んでいる。

ここでは、音楽に例えて、文書が作成・蓄積される過程と時間の関係の検討を行う。

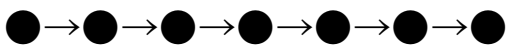
作曲は図 3.1 のように表せる。「●」は作曲過程の結果として他人に提示されるもの、すなわち作曲が完了した曲を表す。「○」は作曲途中の曲であり、曲の旋律の一部やなんらかのモチーフが該当する。図 3.1 の「●」である曲全体が、フッサールのいう第二次記憶に相当する。

これに対し、演奏は図 3.2 のように表せる。演奏者が奏でた音は常に鑑賞者に届き、時間の経



→: 時間経過 ●: 結果

図 3.1 作曲における時間経過と結果の提示



→: 時間経過 ●: 結果

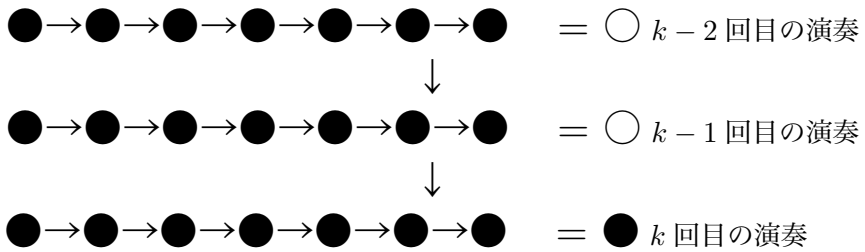
図 3.2 演奏における時間経過と結果の提示

過と共に、曲として聴く人に提示される。

このように、フッサールのいう第二次記憶に当たる作曲は、時間経過に沿った行為の積み重ねだが、作曲そのものは外部に提示される結果ではない。作曲という行為を経て作られた曲が外部に提示される。一方、第一次記憶に相当する演奏は、行為の積み重ねであると同時に、外部へ提示される結果でもある。

3.1.2　演奏の繰り返しによる変化

図 3.2 のみを見ると、演奏者は時間経過に沿った行為の積み重ねを経ずに、いきなり「結果」を提示しているように見える。録音再生を行う機械であればそのように考えることも可能だが、実際の演奏者は、演奏やその練習を繰り返しており、演奏の繰り返しを通して演奏形態を変化させていると考えられる（図 3.3）。すなわち、演奏者は「演奏」を行うまで、どのような演奏を行うか、演奏内容を変化させることができ、「演奏」はその「結果」である。すなわち、時間の経過を貫いて存在する「演奏者」「曲」により、第一次記憶が積み重ねられ第二次記憶を形成している。



→、↓: 時間経過 ●: 結果

図 3.3 演奏者のたどる時間経過と結果の提示

作曲と演奏の中間の形態として、即興演奏、すなわち事前の明確な作曲行為を伴わない曲の演奏がある。即興演奏は、事前の作曲無しとはいえ、音楽の経験のない人間が突然行えるものではなく、それまで音楽経験をふまえ、以前の作曲・演奏や、その演奏中のその時点までの演奏の聴衆への影響などを考慮し、作曲を行いながら演奏を行うものである。すなわち、即興演奏の中には、その時点までの演奏の聴衆への影響などを考慮して行われる短い時間の積み重ねと、場を変えるごとの即興演奏の繰り返しのように相対的に長い時間の積み重ねとの、複合的な積み重ねが含まれている。

3.1.3 作曲の過程

作曲も、1 つの曲を完成させて終了するのではなく、また別の曲を作曲することにつながっているととらえることができる。

また、作曲が完了するまで、曲は「聴衆」に対する演奏はされないが、作曲者自身に対しては、作曲に伴い作成された曲の断片が、思考の中、あるいは手元の楽器を用いて演奏され、第一次記憶に相当する結果となり、曲に新たな要素を加える出発点となる。

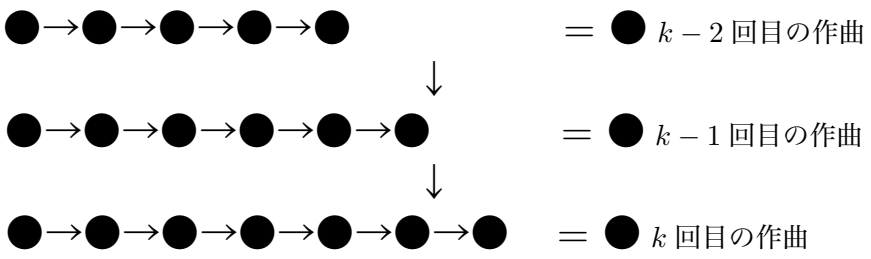


図 3.4 作曲のたどる時間経過と結果（曲）の提示

図 3.4 に、作曲を作曲者自身への演奏を含む過程として捉え直した場合の図を記す。図の等号の左側の「●」は作曲者に対する結果を示し、図の等号の右側の「●」は最終的成果物である曲を聴く人間に対する結果を示す。

3.1.4 本研究で着目する時間の経過

即興演奏、作曲どちらの場合も、時間の経過として、一つ一つの行為をその場で振り返る第一次記憶に相当する時間経過（図中の「→」、以下、短い時間経過）と、それらの積み重ねを振り返る第二次記憶に相当する時間経過（図中の「↓」、以下、長い時間経過）が存在した。

1 つの文書の解析、例えば、深層学習を用いた言語モデルは、1 つの文書（を構成する文章）の

生成過程のモデル化を目的とし、短い時間経過を扱う。これに対し本研究では、長い時間経過、すなわち文書集合内の複数の文書の間に渡る時間の経過（図中の「↓」）に着目する。

3.2　文書と記述対象の関係

続いて、文書に記述される対象の種類と文書集合の蓄積過程の関係について検討する。対象として、2.2.2 節でも述べたニュース記事などのあるタイミングで発生し変化しない出来事と、1.1.2 節で述べた設計過程などの時間経過に伴い変化するものの、2 種類をとりあげる。

3.2.1　出来事を記述した文書集合

出来事は、時間経過のある時点で発生し、発生の前・後に変化することはない。出来事には、あるタイミングで 1 回だけ起こったもの（たとえば単発の事件や事故)、あるいは出来事間の関連性があるもの（たとえば経済動向や政治動向）などがある。ある物に関する記述も「記述者がその物を見た」ことを記しているため、出来事の記述に分類できる。

本節では、あるタイミングで 1 回だけ起きた出来事について記述した文書からなる集合について、検討する。このような文書集合は、過去、場合によっては未来のある出来事について、文書が作成された時点（正確には作成された時点の少し過去の時点）の情報を文書記述者の視点から述べた文書が、逐次的に作成され、それらが蓄積されることにより形成される。報道を行う新聞記事はこの形の文書であり、ある出来事についてあるタイミングごとに（例えば、新聞記事なら新聞発行のタイミングごとに）文書が作成され、蓄積される。

出来事を伝える文書集合の意義は、出来事を文書として記録し、記録を広める、あるいは蓄積することにある。そのため、文書の記述者は、出来事に対して文書の記述だけを行う。出来事に対して何らかの働きかけを行う場合があるとしても、記述した文書を通してであり、直接的に働きかけることはないと考えられる。

図 3.5 に、以上の文書と記述対象の出来事と出来事を観察し文書を記述する作成者の人間の関係を示す。

また、出来事ひとつは、ある時点で 1 回だけ発生したものであるとしたが、その出来事の引き金となった出来事、その出来事に影響されておきた出来事など、出来事が連鎖をなすこともある。この場合、対象間に何らかの関係が存在すると考えることができる。この場合の関係を図 3.6 に示す。

なお、図 3.5、図 3.6 には出来事の発生に何らかの行為により関わりをもつ人間を記したが、自然現象の発生のように、人間が関わらない出来事も存在する。

時間経過に沿ったトピックの抽出を行う研究の多くは、2.2.2 節に例としてあげたように報道記事の集合を扱っており、図 3.6 の出来事間の関連性を文書集合から抽出することを目的とする。

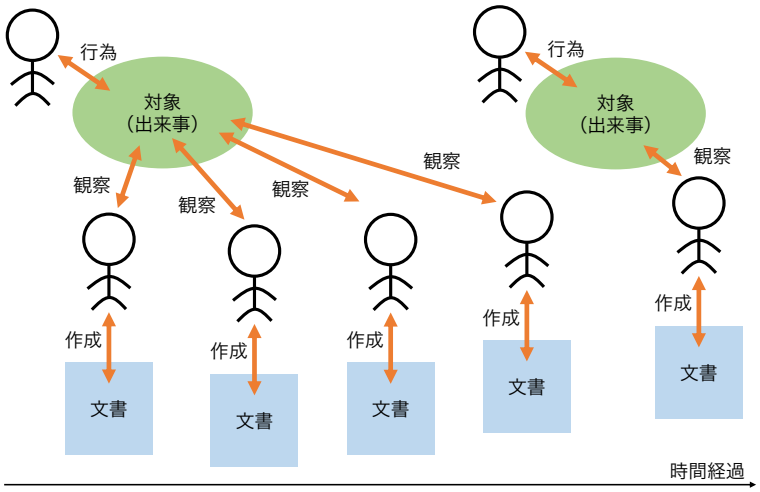


図 3.5 対象が個別の出来事の場合の人間と文書・対象の関係

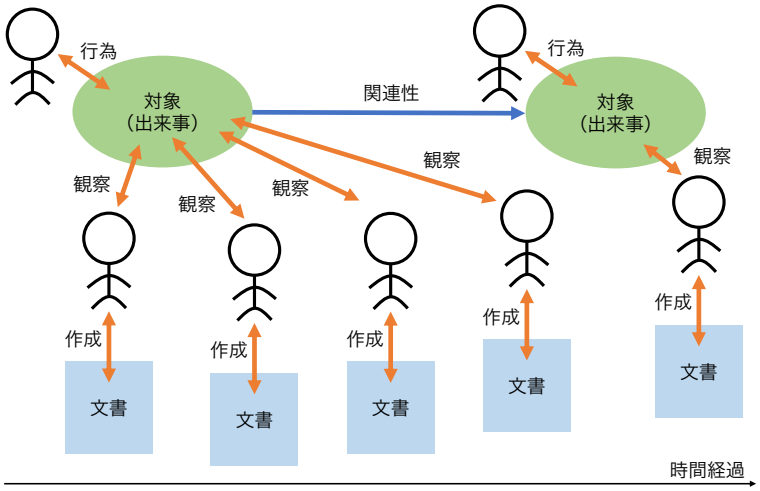


図 3.6 対象が関連するの出来事の場合の人間と文書・対象の関係

3.2.2　時間経過に伴い変化するものを記述した文書集合

次に、時間経過に伴って連続して変化するものを対象とし、時間経過に沿って記述を行った文書の集合を考える。図 3.7 に、時間経過に伴い変化する対象と、対象を観察し文書化する作成者、作成された文書の関係を示す。

時間経過に伴って連続して変化する対象は、1 回限りの出来事と異なり、ある時点では確定したものではなく、文書が蓄積されていく時間過程に沿って変化をしていく。たとえば、政治や経済の動向を継続して追った文書や、会議の過程の記録、植物など成長するものの観察記録が考えられる。その他、文書作成者の考えも時間経過に沿って変化するものである。たとえば、日記は出来事の連鎖について述べているといえるが、時間経過に伴い変化する記述者の主観を述べてい

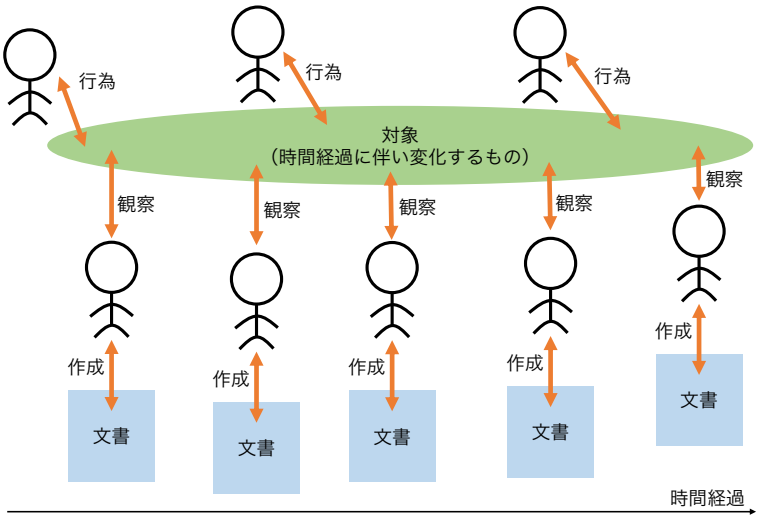


図 3.7 対象が時間経過に伴い変化する場合の人間と文書・対象の関係

ると考えることもできる。

文書の作成は、作成者が文書を記述しようと思い立つところから始まり、文書の変更を完了したところで終わる。この間に対象が変化してしまうこともあるが、ここでは、文書の作成開始から終了までの時間経過は対象の変化に対して無視できるほど短時間であると考えることにする。

これらの文書が作成されて蓄積される意義は、対象の様子を連続して伝達すること、対象の様子を記録し後から変化の過程を参照することの、大きく 2 つに分けることができる。前者の場合、作成者は対象の変化に直接に関わることは、文書の目的からして行われないと考えられる。後者は、対象の変化過程に作成者が関わる場合と、作成者が変化を観察するのみで関わらない場合とがあると考えられる。

3.3　時間経過に沿って変化する対象の文書への記述のされ方

対象が時間経過に沿って変化するものである場合に、3.2 節では、対象は時間経過の中でもひとかたまりの「対象」であるとして考えた。本節では、時間経過に沿って対象を記述する場合、対象の変化の記述のされ方にどのような場合があるのか、考察を行う。

3.3.1　探索木と時間経過

時間経過に沿って変化する対象の例として、問題の解を探索する場合を考える。これには、ゲームなどにおいて次の一手を探す場合が相当する。問題の解を探索する様子は、一般に、複数ある解候補の選択肢を分岐とした探索木として表される。

探索木では、木の分岐点ごとに問題に対する解へ到達する途中の解候補を置き、その次にどち

らの解候補に進むかを、分岐として表す。探索木をたどり解を探索している途中で、解候補が解に到達しないと判断されると、その候補を放棄し、1 つ前の分岐点まで戻り、別の候補から探索を再開する。

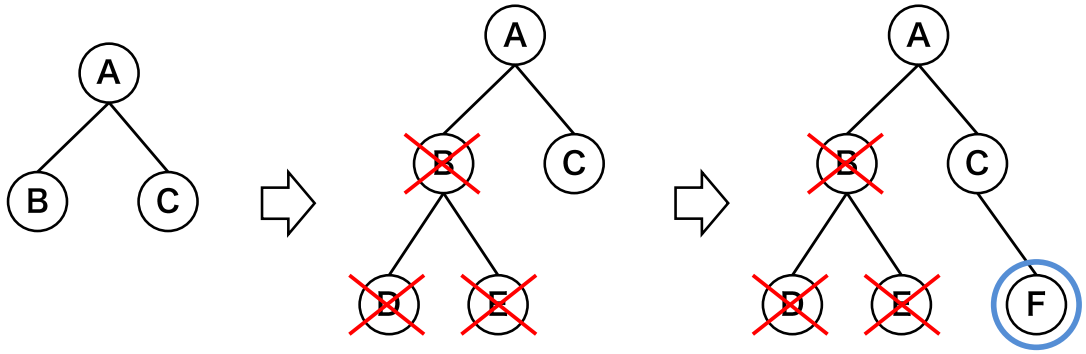


図 3.8 解の探索過程の例

図 3.8 に、探索木により表現された探索の例を示す。A、B、...、F がそれぞれ解の候補を示す。

A から問題を解決するための探索が開始され、F が解である。図 3.8 の例では、途中、D、E が解には到達しないことが何らかの理由がわかったため、A から B（その下位の D、E も含め）へ至る探索過程が放棄される。その後の探索により、A から C をたどり F に到達するルートが、解へたどり着く。

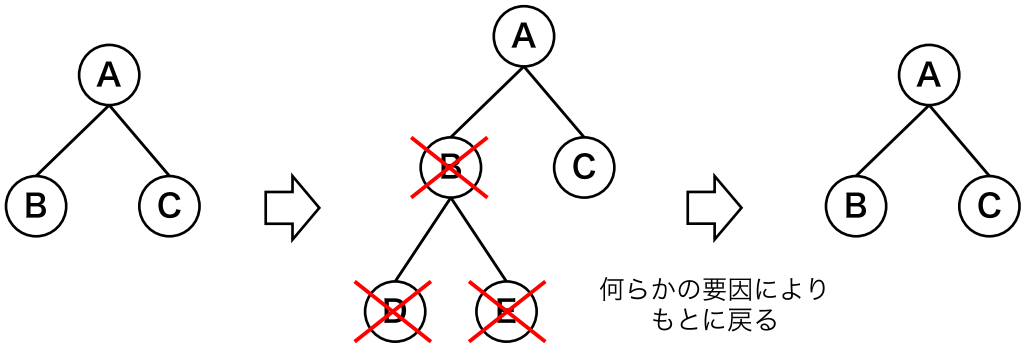


図 3.9 解の探索が進まない例

問題の解決が実時間に伴って進展することが理想だが、現実には、図 3.9 のように、実際の時間は経過するものの問題の解決は進展せず、同じことの繰り返しが発生する場合も考えられる。この状況を記録した文書集合では、文書の作成時刻は進むが、記された内容に変化はない。このような文書集合中では、時間の経過が意味を持たず、文書集合には図 3.9 の任意の部分が時間の経過とは無関係に含まれることになる。

3.3.2　探索木の文書への記述

時間経過に沿った記録として文書を作成するとき、文書における対象の記述方法には表 3.1 の 4 通りが考えられる。

No.	記述方法
1	図 3.8 の最終的な解 F
2	図 3.8 の最終的な解に至る手順 A → C → F
3	図 3.8 の探索過程で放棄された A → B → D なども含むすべて
4	図 3.8・図 3.9 の探索されなかった部分（D の先など）も含む、存在可能性がある探索木全て

表 3.1 解の探索過程の文書への記述方法

これらの情報の記録の分類について、人工物の設計では、対象のある時点での状態を記述した情報、上記 1. や A、B、...、E など各分岐点そのものを記述した情報を「プロダクト情報」、それらに至る過程を表す表 3.1 の 2、3 を「プロセス情報」と呼ぶ [54]。例えば、完成した設計対象を示す設計図はプロダクト情報であり、設計過程で検討された事項に関する情報はプロセス情報と呼ばれる。

出来事も、その出来事が発生するまでには何らかの過程があったはずである。しかし「出来事を記した文書」の記述の中心となるのは、表 3.1 の 1 でありプロダクト情報である。一方、時間経過に沿って変化するものを記述する場合、時間の経過に沿った記述を行うことから、2 または 3 のプロセス情報である。

3.4　人間と文書の関係

3.4.1　人間の行為を記述対象とする文書集合

3.1 節では、曲を文書とみなして、文書と時間の関係を検討した。ここで、曲を図 3.7 の時間経過に沿って変換する対象とみなし、作曲の作業状況や曲の説明などを記した文書を作成し蓄積する状況を想定すると、曲と作曲者、文書の関係は図 3.10 のように考えることができる。

さらにこの状況を、図 3.7 のように複数の行為者がいる状態に一般化すると、図 3.11 となる。

図 3.11 の場合、文書の記述対象となるのは、時間経過に伴い変化する対象へ人間が何らかの行為を行っている状況である。

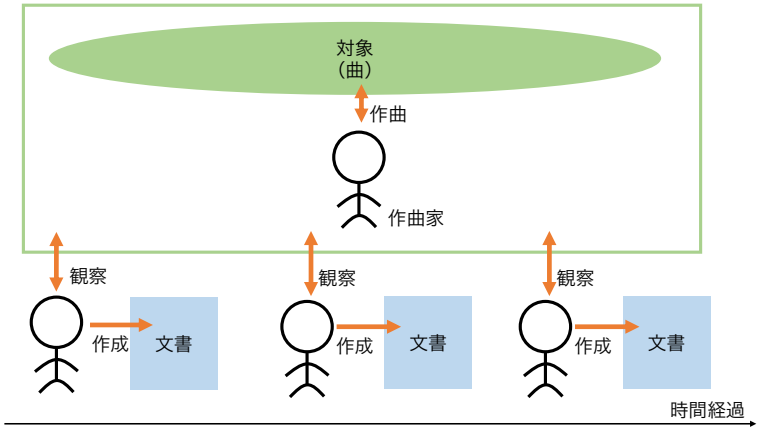


図 3.10 作曲過程を観察した文書を蓄積する場合の関係

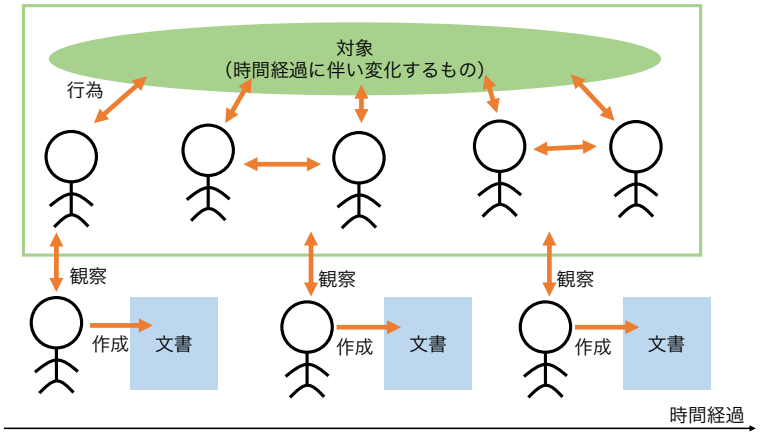


図 3.11 行為の過程を観察した文書を蓄積する場合の関係

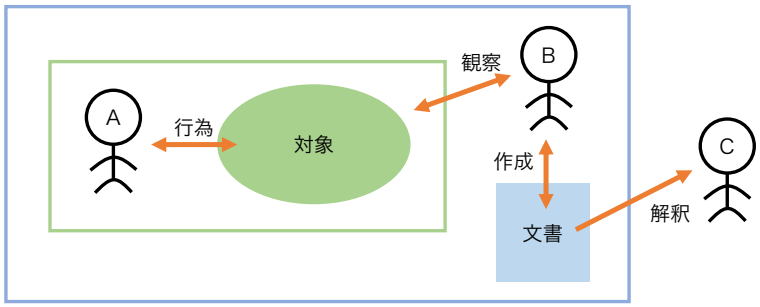


図 3.12 人間・文書・対象の関係の一般化

3.4.2　人間・対象・文書の関係の一般化

図 3.11 に対し、文書を解釈する人間を加えると、図 3.12 のように整理できる。すなわち、文書・文書集合は、対象へ人間 A が働きかけを行う様子を観察した人間 B により作成され、作成された文書・文書集合はそれを解釈し利用する人間 C により活用される。なお、対象は単数でも

複数でも構わず、A と B、C は同一人物でも異なる人物でも、単数でも複数でも構わない。例えば 4 章で提案するシステムの利用者は、人間 C に当たる。

「観察」「作成」「解釈」も「行為」の 1 つであり、「文書」も対象であると考えると、図 3.12 は多階層モデル [34] をなしている。

3.5　時間経過に沿った文書の蓄積のされ方

表 3.2 に、ここまでに述べた時間経過に沿って文書が蓄積される 3 通りの過程をあげる。

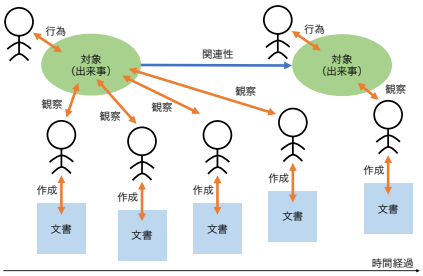
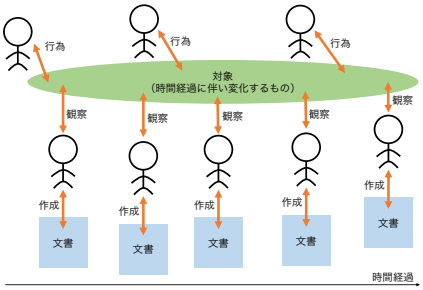
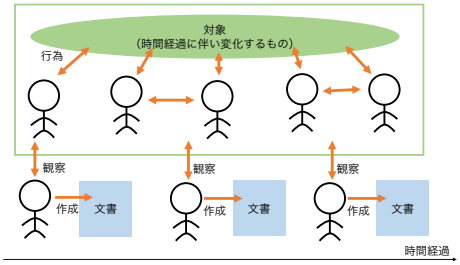
図 3.6		ある時点で発生した出来事を観察し、記述した文書が蓄積される。出来事の間には関連性がありうる。
図 3.7		時間経過に伴い変化する対象を観察し、記述した文書が蓄積される。
図 3.11		時間経過に伴い変化する対象への人間による行為を観察し、記述した文書が、時間経過に沿って蓄積される。

表 3.2 時間経過と対象・文書・人間の関係

図 3.6 の出来事を記述した文書を蓄積する場合、出来事間に関連性があることによって、文書集合内に時間経過に伴った推移が含まれる。関連性がほぼない場合、文書の作成時刻に前後があっても、文書の記述内容は相互に関連せず、文書集合内では実時間の経過が意味をなさない。

図 3.7 の、時間変化する対象を記述した文書を蓄積する場合、対象の変化に伴って、文書の記述内容に時間経過に沿った推移が発生する。しかし、対象の変化がごく小さい場合や、図 3.9 のように、何らかの要因により元に戻る状態が発生している場合、作成時刻の経過が文書集合内で

は意味をなさない可能性がある。

図 3.11 の対象への行為を観察した結果を文書集合として蓄積する場合も、文書の記述対象、すなわち「対象と人間の行為の関係」の変化に伴って、文書の記述内容に時間経過に沿った推移が発生する。「対象と人間の行為の関係」にほとんど変化がない場合、文書集合内では時間経過が意味をなさない。

すなわち、時間の経過に沿って蓄積された文書集合の分析を行う際に、時間の経過が意義を持つためには、文書が何らかの対象について記述を行っており、それらが相互に関連するか、時間の経過に伴い変化するかのいずれかである必要がある。

また、図 3.7、図 3.11 とも、図中には時間経過に伴い変化する対象を 1 つしか記していないが、図 3.6 のように、複数の対象についての記述が文書集合中に含まれる場合もある。

3.6 本章のまとめ

本章ではまず、時間経過に沿って蓄積される文書の集合について、文書集合内の文書の間にまたがる時間の経過に着目することについて述べた。

次に、文書に記述される対象および文書の作成のされ方について検討を行い、対象が時間経過に伴い変化するか否か、変化の過程の全体・あるいは結果のみのどの部分を文書化するのか、また、文書の記述対象として人間の行為を含むか否かにより、文書集合にはさまざな蓄積形態が存在することを整理した。

続いて、文書集合に記述された対象間に関連性がない場合には、文書集合自体は時間経過に沿って蓄積されていたとしても、内容からは時間の経過を推測できない可能性があることを示した。

以上を踏まえ、文書集合から時間経過に沿った内容の移り変わりを抽出しつつ、特定の対象に着目して参照することを可能とするシステムを、4 章にて述べる。

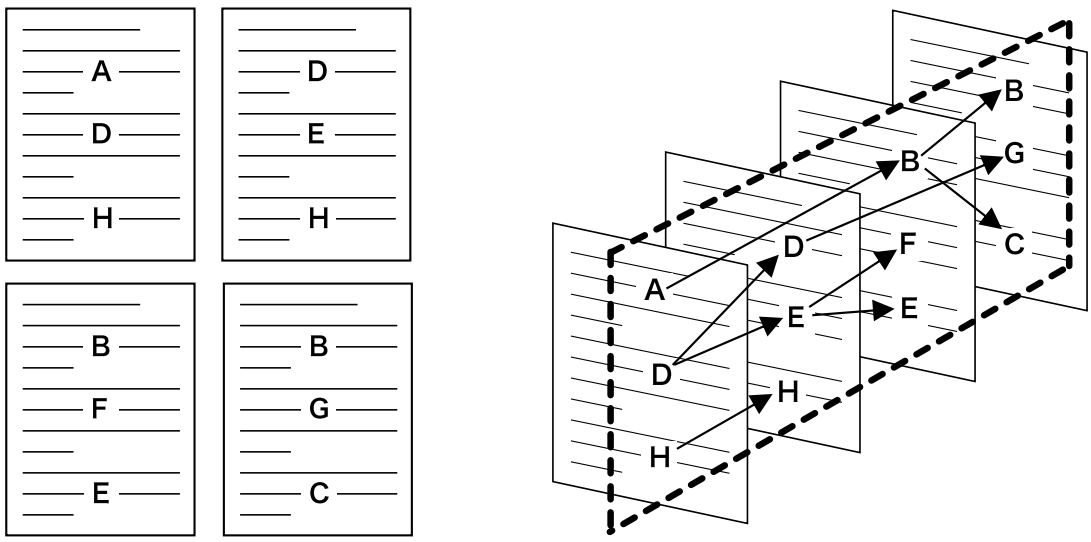
第 4 章

トピック遷移構造の抽出・再構成システム

本章では、時間の経過に沿って蓄積された文章集合から、トピックの遷移を抽出し、対象に合わせて再構成を行うシステムについて述べる。

4.1 トピック遷移構造

時間経過に沿って蓄積された文書集合には、複数のトピックが含まれている（図 4.1(a)）。文書集合全体には、複数のトピックが、時間の流れに沿って混じり合いながら存在し、トピックとその関係からなるグラフ構造をなすと考えられる（図 4.1(b)）。以下、このグラフ構造を、トピック遷移構造と呼ぶことにする。



(a) (b)

図 4.1 文書集合に対し時間経過を意識しない場合 (a) と意識した場合 (b)

4.2　トピック遷移構造の抽出・再構成システムの枠組み

提案するシステムはまず、時間経過に沿って蓄積された文書集合からトピック遷移構造を抽出する。次に、システムの利用者に対し、抽出したトピック遷移構造全体の提示、および利用者の指定単語に基づいたトピック遷移構造の再構成と提示を行う。システム全体の枠組みを図 4.2 に示す。

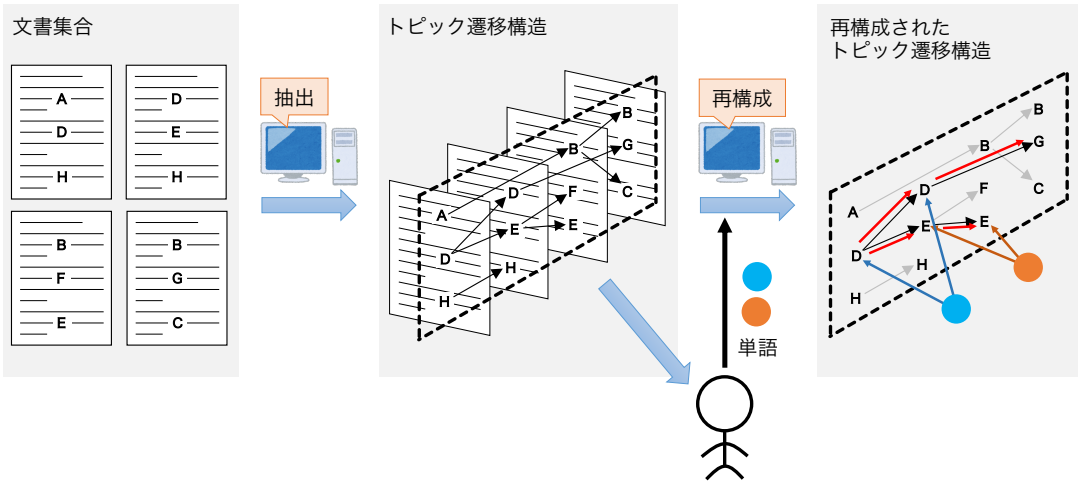


図 4.2 トピック遷移構造の抽出と再構成を行うシステム

本章ではまず、図 4.2 の左側、トピック遷移構造の抽出・表示を行うサブシステムについて述べ、続いて図の右側、トピック遷移構造の再構成と表示を行うサブシステムについて述べる。

4.3　トピック遷移構造の抽出

トピック遷移構造の抽出は、図 4.1(a) のように時系列に沿って存在する複数の文書から、図 4.1(b) のように、時間経過に沿ったトピックと、それらの間の関連を取り出すことにより行う。

さらに、提案するシステムでは、抽出したトピック遷移構造を俯瞰的に提示するだけでなく、利用者の指定に応じてその一部分を再構成して提示する。そのため、トピック遷移構造内には、抽出したトピックごとのキーワードの出現確率などの情報を保持しておく。

なお、2.2.1 節に述べた Dynamic Topic Models などでは、トピックが時間経過全体を通して存在すると仮定してトピックの抽出を行う。一方、本手法では、文書集合内でのトピックの新規発生や消滅を仮定し、また、抽出結果をもとに、トピック遷移構造の再構成など、単語・文書とトピック間の確率を利用した分析を実現するため、独自の抽出手法を用いる。

4.3.1 抽出手法の概要

トピック遷移構造の抽出は、以下の手順により行う。

1. 文書作成時刻による文書集合定義
2. 文書の断片化
3. 文書断片からのトピック抽出
4. 古いトピックの忘却
5. トピック間の関連度計算

1 により時間方向の視点を文書集合に与え、2、3、4 により内容の抽出を行う。1 で定義した文書集合にあわせて 3、4 によりトピック抽出と忘却を繰り返した後、抽出したトピックの時間経過に沿った移り変わりを 5 によりグラフ構造として抽出し、これをトピック遷移構造とする。

続いて、各手順について述べる。

4.3.2 作成時刻による文書集合の分割

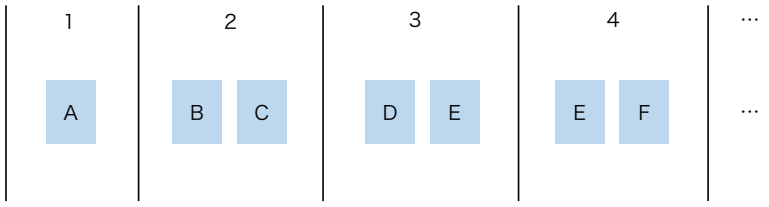
文書集合から時間経過に沿ってトピックの抽出を行うためには、図 4.1(b) のように文書を時系列に沿って並べる必要がある。また、トピックを抽出のためにはトピック抽出の対象となる文書集合を定義する必要がある。そこで、文書の作成時刻により文書集合をいくつかの部分文書集合に分割する。

文書集合の分割方法

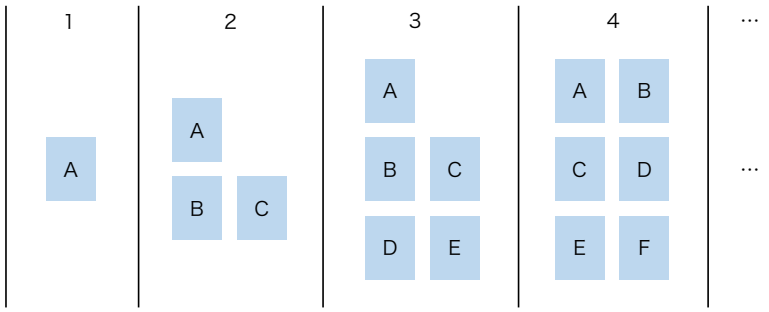
時間経過に沿った部分文書集合の作り方には、図 4.3 に示すようにいくつかの手法が考えられる。図 4.3(a) は一定間隔ごとに区切りとなるタイミングを設け、各タイミングの間に作成された文書を部分集合に分割する方法、(b) は各タイミングまでに生成されたすべての文書を部分集合とする方法、(c) は各タイミングまでに生成された文書を、ある程度の範囲を設けて部分集合とする方法である。

(a) では隣接する文書集合間のトピックに関連が見つけられなかった場合、トピックの遷移がとぎれとぎれになってしまう。(b) では、「古い」トピックがいつまでも残り続けてしまう。望ましいのは (c) のように古いトピックが忘却されていく文書集合である。

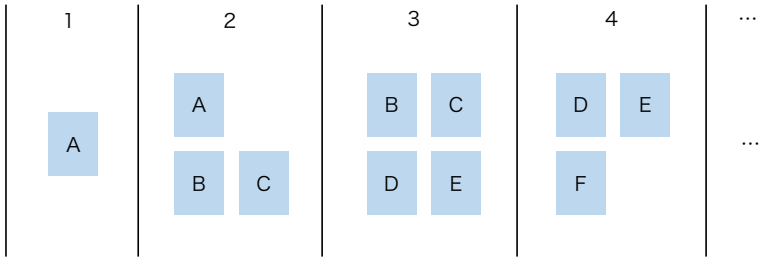
本研究では、一時的に (b) の手法により文書集合を分割し文書集合を定義、その後の処理で古いトピックに属する文書を段階的に忘却させることにより、(c) に類似する文書集合を作成する方針をとる。



(a)



(b)



(c)

図 4.3 文書集合の時刻によるグループ化

部分文書集合の定義

図 4.3(b) のように文書集合から部分文書集合を定義する。

文書集合 D に対して、もっとも古い文書の作成時刻と最新の文書の作成時刻の間を N 等分し、文書集合の分割用時間間隔 S を定義する。$E(D)$ を D 中の最初の文書の作成時刻、$L(D)$ を D 中の最新の文書の作成時刻とし、文書集合の分割時間間隔 S を以下のように定義した。

$$S \equiv \frac{L(D) - E(D)}{N}$$

S に基づき N 個の部分文書集合 $D_1, D_2, \cdots, D_N$ を以下のように定義する。なお、$c(d)$ は文書 d の作成時刻とする。

$$D_n \equiv \{d \mid c(d) \leq E(D) + n \cdot S\}$$

これにより各部分文書集合は、$D_1 \subseteq D_2 \subseteq \cdots \subseteq D_N = D$ となる。

先に述べた通り、この段階では、文書集合は図 4.3(b) に相当する分割を行ったことになるが、4.3.5 にて述べる忘却処理を行い、古い文書の重みを 0 に近づけることにより、図 4.3(c) を実現する。

4.3.3 文書の断片化

ある程度以上の長さを持つ文書は 1 つ以上の話題を含んでいる。例えば、図 4.4 に示す小型人工衛星設計会議の議事録は、1 つの文書の中に、全体の進捗報告、各設計担当ごとの課題、ある特定の話題についての詳細な議論を含んでいる。これらをそれぞれ別の内容へ属するものとして取り扱うために、文書をより短い断片に分割する。

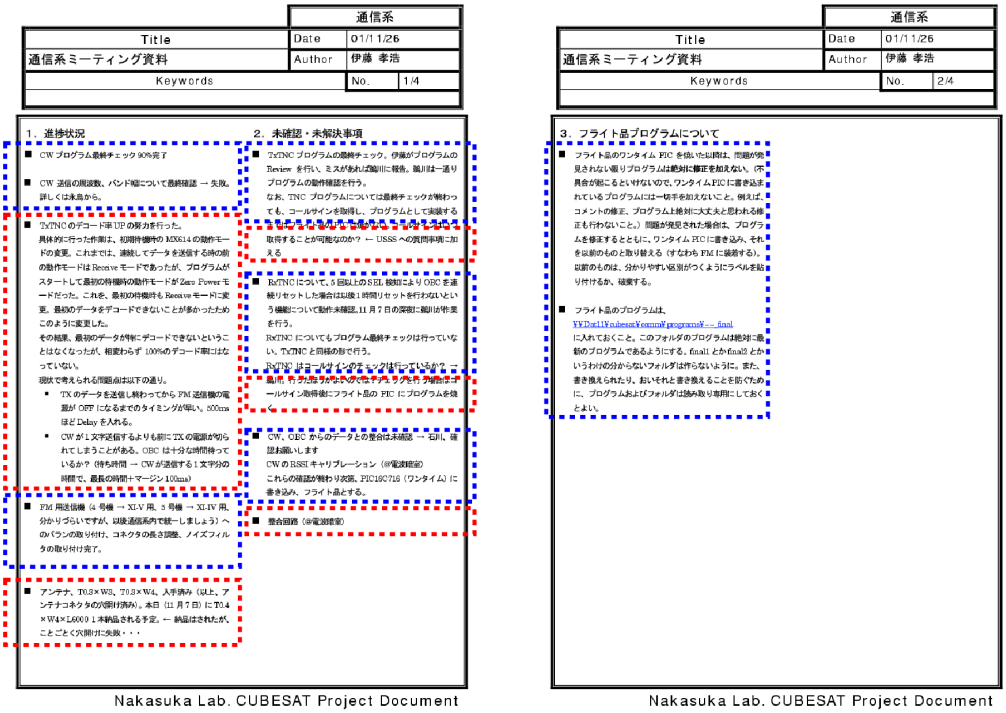


図 4.4 人工衛星設計プロジェクトにおける議事録の例

断片化時に分割点が単語上にある場合は、その単語を文書断片内部に含むよう、W を必要なだけ拡大し、ひとつの単語が分割されてしまわないようにした。また、ひとつの内容が複数の文書断片に分割された場合、重なり部分が同一であることから、トピック抽出時にこれらの文書断片が同一のトピックに分類されることを期待し、連続する文書断片に $\frac{W}{3}$ の重なりを設けた。

文書を断片に分割し、記述されている複数の内容を文書から取り出す手法は、Text Tiling[23]、TDT[6] などにおいて研究されている。

一方、トピック遷移構造の抽出では、断片化に続いてトピックの抽出を行うことから、文書表現にはついて考慮せず、文書をある一定の長さ（W）により分割することで文書の断片化を行

い、断片の集合からトピック抽出を行うことにより、似た断片からなるトピックが抽出されることを狙う。断片の長さと抽出されたトピック遷移構造の性質に関しては、5 章以降にて、文書集合に本手法を適用した結果を確認しながら検討を行う。

4.3.4　文書断片からのトピック抽出

次に、文書集合 D_k から作られた文書断片の集まりを対象としトピックの抽出を行う。トピックの抽出には、トピックモデルの一つである Probabilistic Latent Semantics Indexing (PLSI)[25] を用いる。

各文書断片について形態素解析器によって形態素解析を行い、名詞と品詞分類された単語をとりだし、単語とその出現回数からなる文書ベクトルを作成した。作成した文書ベクトルから M 個のトピックを PLSI により抽出した。形態素解析には Mecab[29] を用い、文書集合に応じて、事前にいくつかの単語を形態素として登録しておいた。

4.3.5　古いトピックの忘却

文書が時間経過に沿って作られていく過程で、以前に記述した内容と類似した記述がなされているとき、その内容は文書集合作成者たちの着目対象であり続けると考えられる。一方、類似した記述がなくなれば、その内容は忘れられつつあるものであると考えられる。

そこで、D_{n+1} に属する文書断片からのトピック抽出後、D_n には存在せず D_{n+1} には存在する文書断片（すなわち新たな断片）を「含まない」トピック $C_{n+1,i}$ に属する文書断片を、忘れ去る対象として重みを減らすことにした。これは、D_{n+2} 以降の処理に用いる文書ベクトルを作成する際、単語の重みを R 倍 $(R < 1)$ することにより実施した。これにより、図 4.3 の (c) に準じた文書集合の設定を行える。

新規文書断片を含む・含まないの判断

新たな文書断片を含む・含まないの判断は、PLSI により得られるトピック z_i に対する文書 d_k の出現確率 $p(d_k|z_i)$ を用いる。$c(d_k) \geq (n-1) \cdot S$ である文書 d_k がトピック z_i に含まれるかを判断するためには、以下のような方法が考えられる。

1. z_i における d の出現確率が他のすべてのトピックよりも大きい
2. $p(d_k|z_i)$ があらかじめ定めたしきい値以上である
3. $p(d|z_i)$ が大きい、すなわち z_i への寄与率が高い d から順に $p(d|z_i)$ の和を取り、しきい値以下の範囲に d_k が含まれている

本システムでは、2 の手法は、文書集合によって $p(d_k|z_i)$ の分布が異なり適切なしきい値の設

定が難しいことから用いず、1 の出現確率が最大となるトピックに文書 d_k が属していると判断する手法、3 のトピックについて一定の寄与率の範囲に d_k が含まれているかを判断する手法を用いた。

また、排他的クラスタリングを行い、その結果、$c(d_k) \geq (n-1) \cdot S$ である文書 d_k がトピック z_i に該当するクラスタに含まれる場合に新規文書と判断する手法も併用した。

トピックの忘却処理

PLSI によるトピック抽出では、処理対象の文書を含まれる単語とその出現回数からなるベクトルとして扱う。文書の重みの操作は、ベクトルの要素である単語の出現回数の操作により行った。すなわち、文書の重みを R 倍する際には、文書を示すベクトル中の単語の出現回数をすべて R 倍した。

4.3.6 トピック間の関連度計算

隣接する文書集合 D_n, D_{n+1} に属するトピック $C_{n,i}, C_{n+1,j}$ 間の関連度として、$sim(C_{n,i}, C_{n+1,j})$ を以下のように定義した。

$$sim(C_{n,i}, C_{n+1,j}) = \frac{|C_{n,i} \cap C_{n+1,j}|}{|C_{n,i}|}$$

$C_{n,i}$ は文書集合 D_n の i 番目のトピック、$|C_{n,i}|$ は $C_{n,i}$ に属する文書断片の数を表す。一般に、集合間の関連度を求めるためには、以下の Jaccard 係数が用いられる。

$$Jaccard(C_{n,i}, C_{n+1,j}) = \frac{|C_{n,i} \cap C_{n+1,j}|}{|C_{n,i} \cup C_{n+1,j}|}$$

しかし、$C_{n,i} \subseteq C_{n+1,j}$ となるとき、すなわち、$C_{n,i}$ が示す話題が他のもっと広い範囲の話題をもつ $C_{n+1,j}$ に統合されたとき、Jaccard 係数の分母の値が大きくなり、相関値が小さくなってしまう。そこで、上記の $sim(C_{n,i}, C_{n+1,j})$ を相関関数として用いた。

以上のように、トピックの抽出と忘却処理、トピック間の類似度の似度の計算により、トピック遷移構造を抽出した。

4.3.7 変更可能なパラメータ

トピック遷移構造の抽出を行う手順において、ここまでに述べた変更が可能なパラメータを表 4.1 に示す。

設定	設定値
時間方向の文書集合数	N
文書断片の長さ	W
トピック数	M
トピック忘却時の重み	R
トピック忘却時の新規文書判定手法	確率合計 or 排他的クラスタリング
トピック間のリンクしきい値	T

表 4.1 トピック遷移構造の抽出におけるパラメータ

4.4　トピック遷移構造の俯瞰提示

続いて、抽出したトピック遷移構造を表示するシステムについて述べる。まず、トピック遷移構造全体を俯瞰提示するシステムについて述べる。

俯瞰表示を行うもっとも簡単な手法は、トピックの特徴語を一覧として並べることである。例えば、Blei らは Dynamic Topic Models を用いて抽出された時系列に沿ったトピックを図 2.3 のように記している [7]。

トピック遷移構造は、図 4.1(b) のようにトピック間のリンクも抽出してグラフ構造を得ることから、トピックとそのリンクからなるグラフを表示する。以下、トピック遷移構造の提示システム上に実装した表示方法について述べる。

4.4.1　静的なグラフ提示

俯瞰表示方法の 1 つ目として、トピック遷移構造を、レイアウトを固定したグラフとして提示する方法を設けた。グラウのレイアウトには、グラフレイアウト・可視化ツールの Graphviz[15] を用い、PDF としてグラフを生成させ、Web ブラウザ上の操作用インタフェースを通して表示を行う。例を図 4.5 に示す。

グラフの生成にあたり、横軸に左から右へ過去から未来方向に時間の経過を指定し、縦軸は特に指定を行わず、Graphviz にレイアウトを行わせた。グラフのノードに当たるトピックのラベルには、PLSI にて求めた各トピックにおける出現確率 $p(w|z)$ が大きい語を設定した。グラフを拡大表示することにより確認できる（図 4.6)。また、グラフのラベルにはトピックを一意に識別するキーを含ませ、4.4.4 に述べるトピックの詳細表示を行う際には、このキーを入力することにより、表示したいトピックを指定することとした。

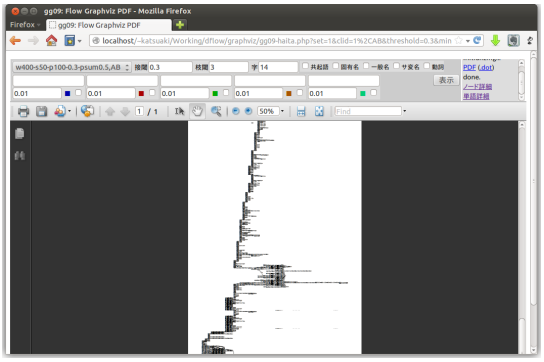


図 4.5 静的なグラフ表示例

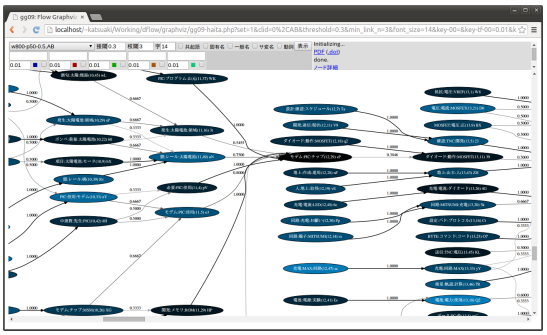


図 4.6 静的なグラフ表示の拡大例

4.4.2　時間軸を固定した提示

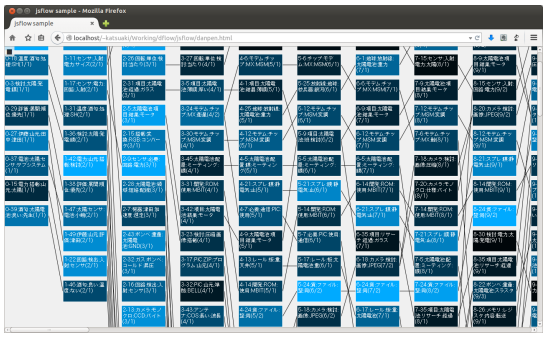


図 4.7 時間軸固定表示

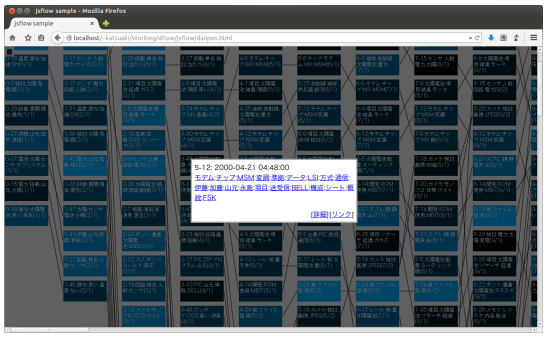


図 4.8 トピック概要表示事

静的なグラフ提示は PDF ファイルにより行ったため、ユーザが指し示したトピックの詳細を表示するなどの操作を行いづらい。そこで、クリックなどの操作を可能にすること、タブレット端末での表示へ対応することなどを目的とし、図 4.7 のようなインタラクティブな操作が可能なトピック遷移構造の提示を行うシステムを、時間軸固定提示システムとして用意した。横軸の左から右が過去から未来方向の時間の経過を示す。縦方向は、トピック間のリンクにより表示時に位置を決定した。

画面上に表示された各トピックを選択することにより、トピックの概要を図 4.8 のように表示し、トピックの詳細や単語の詳細を参照することができる。また、グラフの描画サイズ設定などの操作も行えるようにした。

4.4.3　力学モデルによる提示

トピックの画面上への配置を比較的自由に行う提示システムとして、引力・斥力を用いた力学モデルによるグラフ表示 [18] を行う Arbor.js[10] を用い、力学モデルによる提示システムを設けた。例を図 4.9 に示す。トピック間の類似度によるリンク以外に、隣接する部分文書集合の類似するトピック間にリンクを設ける、4.5.1 節で述べる再構成のキーワードを示すノードを追加するなど、時間の経過に沿ったレイアウトにするための操作を行えるようにした。

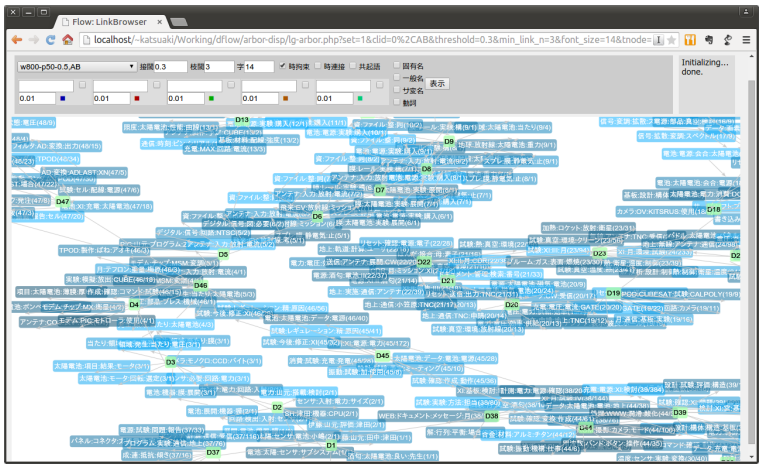


図 4.9 力学モデルによる表示

表示されたノードをダブルクリックすることにより、画面上部のメッセージ領域にそのノードに相当するトピックの概要などが表示される。また、グラフの拡大縮小、選択したノードの表示位置の固定なども行えるようにした。

4.4.4　トピック詳細の提示

各俯瞰表示では、トピック遷移構造のノード、すなわちトピックの詳細を表示する機能を設けた。図 4.10 に例を示す。この画面では、トピックに関連する文書断片と単語の表示を行う。文書断片は、トピックにおける文書断片の生起確率 $p(d|z)$ か、文書断片の元となった文書の作成時刻が新しい順による表示、単語はトピックでの生起確率 $p(w|z)$ が大きい順による表示を行う。

また、文書断片ごとに、文書断片の元となった文書本文を参照するためのリンク、文書断片を含むトピック遷移構造のグラフを参照するためのリンクを示し、これらを表示することができるようにした。元となった文書の表示例を図 4.11 に示す。

図 4.10 トピック詳細の表示例

図 4.11 元文書の表示

4.4.5 トピック遷移の 2 次元配置アニメーション提示

各文書集合の特徴を把握するために、トピック遷移構造中の全トピック $z_{ik}(1 \leq i \leq N, 1 \leq k \leq M)$（$N$、$M$ は表 4.1 参照）を 2 次元平面上に配置し、時間の経過を示す i に沿って z_{ik} のみを表示することにより、時間経過に沿ったトピックの遷移を 2 次元配置アニメーションとして提示する仕組みを設けた。

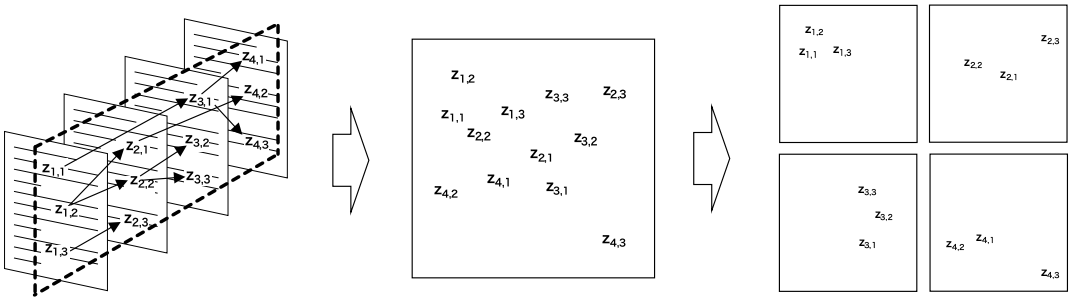


図 4.12 トピック遷移構造からの 2 次元配置アニメーションの生成

トピック遷移構造の抽出により得られたトピックは、抽出に pLSI を用いたことから、D_i から抽出したトピック $z_{ik}(1 \leq i \leq N, 1 \leq k \leq M)$ について、文書中に出現する単語 w の生起確率の分布 $p(w|z_{ik})$ が計算されている。w の意味が文書集合が蓄積された時間経過を通して変化しないと仮定すると、$p(w|z_{ik})$ は、各トピックの文書集合全体における時間経過に依存しない特徴を示すベクトルである。これらを可視化に適した次元削減を行う t-SNE[43] を用いて 2 次元に変換し、プロットすることにより、2 次元表示を生成する（図 4.12）。トピックを示すラベルには、トピック中の出現回数上位 3 語を選択した。

生成されたアニメーションの表示例を図 4.13 に示す。

全体 $i = 10$ $i = 20$ $i = 30$ $i = 40$ $i = 50$

図 4.13 トピック遷移構造のアニメーション表示例

4.5 単語を中心としたトピック遷移構造の再構成と提示

4.5.1 キーワードによるトピック遷移構造の再構成

トピック遷移構造の俯瞰提示に対し、提示システムの利用者が、着眼点として何らかの単語とトピックが単語に関連する・しないを判断するしきい値を指定し、その単語に関連するトピックとリンクをトピック遷移構造から抽出し、提示する仕組みを設けた。以下、システム利用者が指定した単語をキーワードと呼ぶことにする。

図 4.14 に、キーワードによるトピック遷移構造再構成のイメージを示す。利用者によるキーワードの設定、キーワードと関連しているとするしきい値の設定は、各トピック遷移構造の俯瞰表示システムから行えるようにした。力学モデルによる表示システムにおいて、トピック遷移構造の再構成し表示した例を、図 4.15 に示す。

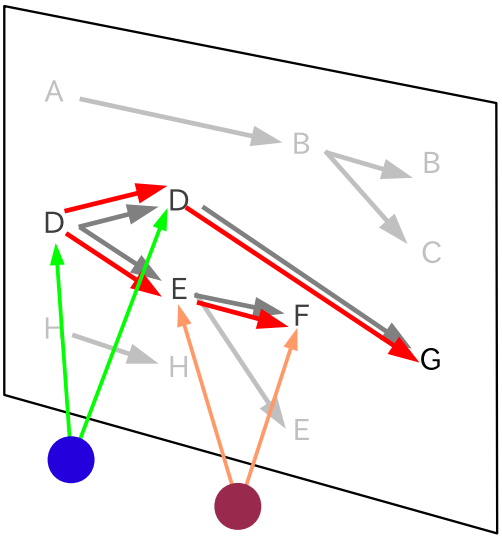


図 4.14 キーワードによるトピック遷移構造の再構成

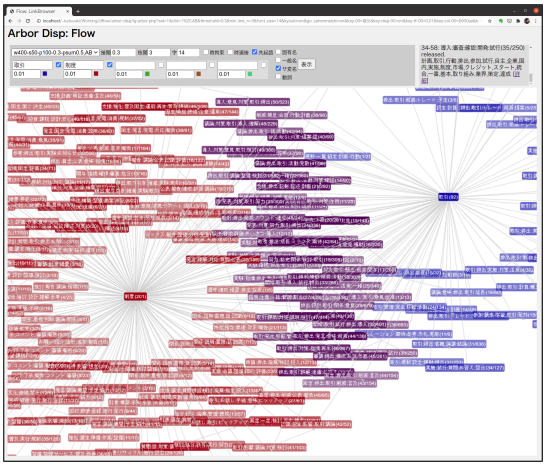


図 4.15 再構成表示例（力学モデルによる表示例）

キーワードとトピックが関連するかの判断は、トピックにおけるキーワードの生起確率 $p(w|z)$ が指定されたしきい値以上であるかにより行う。

また、再構成を行うためのキーワードは複数指定可能とした。これにより、トピック遷移構造の再構成と表示を通し、利用者が新たに興味を惹かれた単語があれば、再構成を行うキーワードを追加することにより、複数のキーワードにまたがるトピックの移り変わりを確認することができる。

4.5.2 再構成キーワードの設定支援

キーワードの設定は、システムの利用者が行う。利用者がキーワードを思いつくままに自由に記述してもよいが、文書集合に出現する単語でなければ、文書集合から得られたトピック遷移構造を再構成することはできない。そこで、キーワードの選択を支援するいくつかの手法を用意した。

まず、トピック遷移構造の俯瞰表示において、ノードのラベルとして利用者へ提示したトピックの代表語が、キーワードの候補となりうる。ノードラベルは各トピックにおいて出現確率 $p(w|z)$ が高い語であり、俯瞰表示を確認することで、複数のトピックにわたって出現確率が高い語をキーワードの候補とすることができる。

その他、候補となる単語の一覧として、図 4.10 に挙げたトピック詳細の表示において、トピック内の文書断片の詳細のみではなく、トピックにおける生起確率 $p(w|z)$ が高い順に単語を表示する機能を設けた。

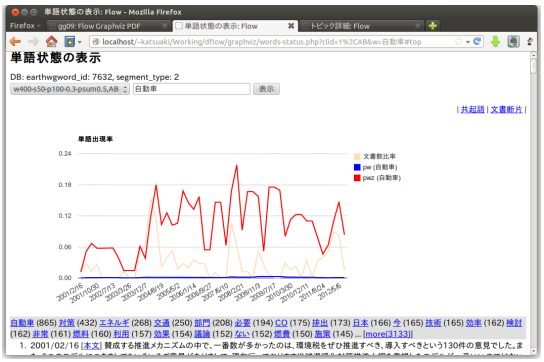


図 4.16 単語生起確率推移の表示例

図 4.17 キーワードのサジェスト例

また、トピック遷移構造の中での単語の生起確率の推移を表示する機能を設けた（図 4.16）。この表示より、興味を持った単語について、時間経過に沿った単語の生起確率を確認し、キーワードと対で設定するしきい値の目安とすることができる。

その他、キーワードの入力時には、図 4.17 に示すように、キー入力に従って、文書集合に含まれる単語をキーワード候補としてサジェストする機能をもたせた。これにより、文書集合に存在しない単語を入力して空振りに終わることをなくし、存在する単語のみを対象として絞り込むことができる。

4.5.3　ラベル語の品詞指定

ここまでに述べたトピック遷移構造の表示システムでは、トピック遷移構造をグラフ表示する際のノード、すなわちトピックのラベルとして、トピック中で生起確率 $p(w|z)$ が高い単語を用いた。

これ以外に、キーワードと同じ文書断片に含まれる単語、すなわち共起する単語を選択して表示する機能を持たせた。あわせて、文書断片から文書ベクトルを作成するために形態素解析を行った際に、形態素解析器が付与した品詞情報を用いることで、利用者が指定した品詞の単語をラベルとして表示することができるようにした。すなわち、トピックに含まれる文書断片から、キーワードを含む文書断片を選び、そこに含まれる指定品詞の語を、ラベルとして選択する。この際、出現時刻が最も新しい文書断片から順に語の選択を行うこととした。

これにより、たとえば、キーワードと共起する固有名詞を指定すると、キーワードと同時にどのような事物が文書に記されているかをグラフとして表示することができる。また、サ変名詞(「〜する」という接続が可能な名詞）を指定すると、キーワードに対して行われたことの連鎖を抽出できる。

これにより、キーワードによる再構成とあわせ、キーワードを中心にどのような話題が存在するかをトピック遷移構造から抽出し、表示するを可能とした。

4.6　本章のまとめ

本章では、文書集合からそこに含まれるトピックの遷移を抽出し、提示を行うシステムについて述べた。

トピックの遷移を抽出するために、まず、蓄積された文書をそれぞれの生成時刻に基づいていくつかの文書集合に分けておき、各文書をより細かな文書断片に分割、文書集合ごとに文書断片から PLSI によりトピック抽出を行い、トピックとそれに対する文書断片および単語の生起確率を得た。続いて、生成時刻が隣接する文書集合から生成されたトピック間の関連度を求め、トピック間のリンクを設定し、時間軸に沿ったトピックとそれらのリンクからなるグラフ構造を、「トピック遷移構造」として得る。これらの機能をトピック遷移構造の抽出サブシステムとした。

また、得られたトピック遷移構造を俯瞰的に提示するシステム、利用者の指定する単語などに基づいて再構成を行い提示を行うシステムについて、それぞれの機能を述べた。

第 5 章

設計議事録の分析

本章では、4 章に述べたトピック遷移構造の抽出・再構成システムを実際のある文書集合に適用し、文書集合に含まれる情報の分析を行う。また、分析を通し、提案システムの各機能が期待した目的に沿った働きを行うことを確認する。

5.1 実文書集合への提案システム適用と目的

4 章で述べたトピック遷移構造の抽出と再構成を行うシステムを、東京大学大学院工学系研究科中須賀研究室により行われた小型人工衛星（CubeSat XI-IV）プロジェクト [58][49] の設計会議議事録など、2000 年 1 月 5 日から 2002 年 12 月 12 日の約 2 年分からなる文書集合に対し適用する。この文書集合は「小型人工衛星」という具体的な対象を共有し、その設計に関わった複数の人間により記述された活動記録である。

本文書集合の記述の主たる対象は「小型人工衛星 CubeSat XI-IV」であり、目的は「小型人工衛星を製作すること」である。「小型人工衛星」はひとつの具体的な対象だが、その中には、人工衛星を構成する複数の対象や、それぞれの対象に関する多くの目的がからみ合っている。そこで本章では、提案システムにより、これら絡み合った内容を俯瞰すること、および、読み手の興味に沿い、内容をときほぐしつつ読み解くことができるかを確認する。すなわち、本論文で提案するシステムにより、本文書集合から、小型人工衛星を製作するために行われた、衛星を構成する部分に対する個別の設計活動内容などを抽出し、その内容や理由を読み取れることを確認する。

5.1.1 対象文書集合の概要

本章で扱う文書集合は、2000 年 1 月 5 日から 2002 年 12 月 12 日までに作成された、CubeSat XI-IV に関する議事録、マニュアル、実験記録など、398 文書からなる。各文書は、作成日付、タイトル、記録者などのヘッダ部、および議事内容を含み、日付、タイトル、記録者以外の部分には統一された書式はなく、記録者によりまちまちの形式で記述されている。また、議事録は、

図 4.4 に示したように 1 回の会議の記録を 1 つの文書にまとめたものであり、1 つ文書は 1 回の会議で検討された複数の内容を含んでいる。

なお、XI-IV を開発した中須賀研究室では XI-IV 以後も多くの人工衛星の設計を行なっているが、XI-IV は一連の活動の初期に製作された衛星であり試行錯誤を多く含む。このため、本章の目的である「複数に絡み合った対象などをほぐして読み解く」ことに適すると考え、XI-IV に関する文書集合を本章の分析対象とした。

5.2　トピック遷移構造の抽出

5.2.1　前処理

CubeSat XI-IV に関する文書は PDF 形式で保存されていた。そこでまず、テキスト形式に変換した。この際、図 4.4 に示したように文書が 2 段組に整形されていることにより、正常にテキスト形式へ変換できなかったものがある。変換後の平均文書長は 1,352 文字であった。

次に、人名や部品名称など複数の単語について形態素解析器の辞書へ登録し、形態素解析を行った。トピック遷移構造の抽出結果を確認し、ひとつの単語として処理することが適切と思われる単語約 100 語を、形態素解析器の辞書に登録した。

5.2.2　抽出結果

続いて、4 章にて述べた手法により、文書集合からトピック遷移構造の抽出を行った。

トピックの代表語

抽出結果の一部のうち、トピック遷移構造に含まれるトピックの代表語の一部を表 5.1 に示す。各単語は、トピック遷移構造の抽出を表 5.2 の設定により行い、各 $C_{n,i}$ について $p(C_{n,i})$ が大きい上位 10 個のトピックについて、$p(w_k|C_{n,i})$ が最大となる単語 w_k、すなわち各時間区間において生起確率が大きいトピックにおける特徴的な単語である。

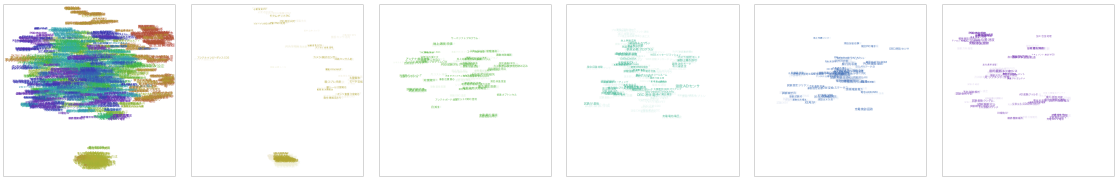
トピックの 2 次元表示

図 5.1、図 5.2 に、5.1.1 節で述べた小型人工衛星の設計議事録に対し、表 5.2 および表 5.3 の設定で抽出したトピック遷移構造を、4.4.5 で述べた手法によりアニメーション表示させたうち、全体、$i = 10, 20, 30, 40, 50$ の状態を示す。

トピックのレイアウトされた位置から、時間の経過に沿って、出現するトピックが変化していることを確認できる。また、断片長 800bytes（図 5.1）、400bytes（図 5.2）のどちらでも、ト

D_n	特徴語									
2	SH	太陽	センサ	温度	センサ	検出	電力	評価	酒匂	伊藤
4	アンテナ	アンテナ	信号	発生	PIC	ボンベ	モデム	アンテナ	太陽電池	電池
6	発生	太陽	飛来	アンテナ	デジタル	整	項目	必	スプレ	PIC
8	電池	電池	発生	膜	飛来	膜	アンテナ	電池	機器	資
10	電池	電池	電池	電池	電池	電池	電池	電池	電池	電池
12	電池	電池	電池	電池	電池	電池	電池	電池	電池	電池
14	BYTE	衛星	通信	電池	電圧	PIN	地上	試験	TNC	構造
16	リンク	データ	動作	リセット	アンテナ	活動	試験	設計	通信	受信
18	書き込み	リセット	サーバ	カメラ	インターネット	太陽電池	RADIO	電池	アンテナ	電圧
20	地上	電源	試験	電力	試験	アンテナ	送信	充電	XI	トルク
22	リンク	基板	地上	電源	試験	リセット	電源	XI	ネットワーク	送信
24	地上	試験	振動	月	無線	カメラ	酒匂	軌道	試験	リセット
26	振動	BIT	温度	XI	試験	無線	月	充電	軌道	温度
28	周波数	XI	実施	試験	BIT	電池	POD	試験	充電	試験
30	温度	OBC	受信	確認	無線	試験	軌道	確認	確認	撮影
32	試験	月	データ	実験	送信	電圧	カメラ	電流	制御	アンテナ
34	試験	ROM	データ	確認	アンテナ	実験	接着	全体	地上	電圧
36	時刻	電池	試験	カメラ	基板	進捗	接着	試験	電圧	ROM
38	XI	実験	試験	XI	XI	作業	解	電源	設計	自分
40	太陽電池	XI	XI	試験	ROM	回路	充電	解	CW	XI
42	試験	地上	調査	XI	ISSL	クリーンルーム	試験	太陽電池	電源	OSSS
44	試験	太陽電池	デル	合金	アンテナ	ロケット	表面	技術	振動	設計
46	アオキ	コマンド	試験	試験	部品	蝶番	処理	月	消費	合金
48	試験	最終	電力	充電	設計	月	AD	消費	運用	試験
50	放出	最終	実験	月	蓋	ロケット	試験	DUTY	電力	ねじ

表 5.1 抽出したトピック遷移構造にて $p(w|z)$ が最大の単語（n = 偶数のみ、上位 10 トピック）



全体　$i = 10$　$i = 20$　$i = 30$　$i = 40$　$i = 50$

図 5.1 小型人工衛星のトピック 2 次元配置（設定は表 5.2）

全体　$i = 10$　$i = 20$　$i = 30$　$i = 40$　$i = 50$

図 5.2 小型人工衛星のトピック 2 次元配置（設定は表 5.3）

ピックの 2 次元表示に大きな違いがないことが確認できる。

5.3　トピック遷移構造の俯瞰表示による分析

抽出したトピック遷移構造を、4.4 にて述べた俯瞰表示により確認する。俯瞰表示は、時間経過に沿ったトピックの集合とトピック間のリンクからなるトピック遷移構造を、ほぼそのまま表示する手法である。

まず、抽出されたトピック遷移構造に含まれるデータを正常に表示できることを、各俯瞰表示手法において確認した。静的なグラフ表示（4.4.1）、時間軸を固定した表示（4.4.2）、力学モデルによる表示（4.4.3）において、トピックにおける生起確率が高い単語などの表示を確認できた。

設定	設定値
時間方向の文書集合数	50
文書断片の長さ	800bytes
トピック数	50
トピック忘却時の重み	0.3
トピック忘却時の新規文書判定手法	断片出現確率合計・0.5 以下

表 5.2 表 5.1・図 5.1 抽出用設定

設定項目	設定値
時間方向の文書集合数	50
文書断片の長さ	400bytes
トピック数	50
トピック忘却時の重み	0.3
トピック忘却時の新規文書判定手法	断片出現確率合計・0.5 以下

表 5.3 図 5.2 抽出用設定

また、4.4.4 に述べたトピック詳細の表示機能により、トピック内で生起確率が高い文書断片と単語を確認することもできた。

俯瞰表示により、分析対象とした文書集合の全期間のトピックを表示させると、例えば図 4.5 のように、非常に細かい表示となる。この表示の一部を拡大すると、例えば図 5.3 ように、トピックの代表語などを確認できた。このように一部を拡大しつつ表示することにより、表 5.1 に繰り返し現れる「電池」「アンテナ」のように繰り返し長期にわたり出現するトピックのみでなく、「モデム」といった短い期間のみ継続しているトピックがあることや、トピック間のリンクを示す線により、トピックが分岐したりまとまったりして絡み合う様子を確認することができた。

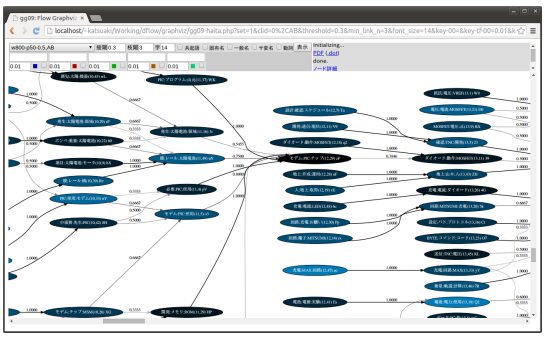


図 5.3 静的なグラフ表示の拡大例（図 4.6 再掲）

図 5.4 時間軸を固定した表示の図 5.3 付近

俯瞰表示では、表 5.1 に現れる単語のように繰り返し現れる単語を含んだトピックは目につきやすい。一方、システムの利用者がトピックの詳細を確認した結果、新たに興味を惹かれた内容・部分があっても、ラベル語に現れない部分は確認することが難しかった。例えば、利用者が

「モデム」に興味を持ったとしても、モデムに関連するトピックを俯瞰表示されたラベル語から、目視で見つけ出すことは困難であった。

5.4　トピック遷移構造の再構成による分析

次に、4.5 に述べた単語によるトピック遷移構造の再構成機能について確認を行う。トピック遷移構造をシステム利用者が指定した単語に基づいて再構成し表示し、トピック遷移構造の部分的な把握が行えることを確認する。俯瞰表示では把握しづらかった、利用者の把握したい部分の確認を試みる。

対象とする文書集合は、すでに述べたとおり、小型人工衛星という構造を持った対象に関する設計記録である。そのため、多くの部品の名称が文書中に出現する。また、本文書集合を読み解く人間も、これらの部品がどのような位置づけにあるかに興味を持つ可能性が高いと考えられる。そこで、トピック遷移構造の再構成を行うキーワードとして、このような部品名称を用い、動作の確認を行う。

以下の確認には、特に記述しない限り、表 5.2 の設定により抽出したトピック遷移構造を用いた。

5.4.1　ひとつのキーワードによる再構成

5.3 でも取り上げた「モデム」について確認を行う。「もで」と途中までキーワードを入力すると、図 5.5 のように、4.5.2 に述べた単語入力時のサジェスト表示機能により、候補となる単語が表示される。この機能により、部品名称が正確に把握できていないなど、指定したい単語の表記があやふやな場合などに、文書集合から得られた単語が表示されることにより、文書中の記述と一致する単語を、表記の間違いなく指定できる。

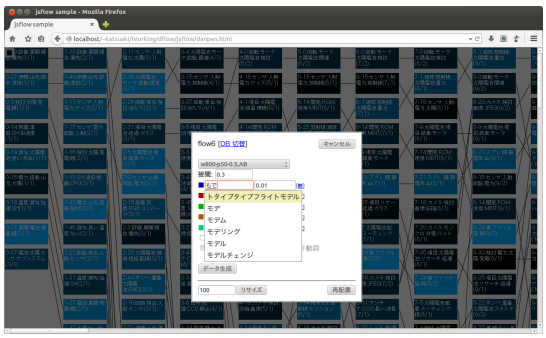


図 5.5 「モデム」入力中の単語サジェスト

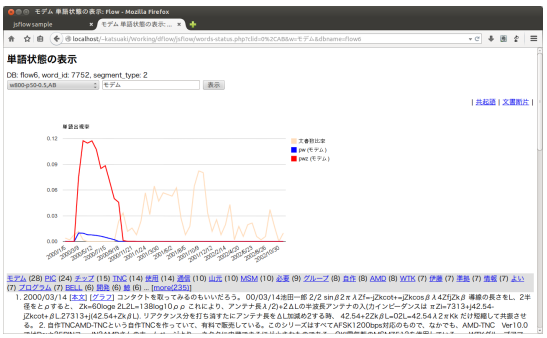


図 5.6 「モデム」の生起確率推移

図 5.6 に、「モデム」についてトピック遷移構造の中での単語の生起確率遷移を表示した結果を

示す。この図より、「モデム」は設計の初期に現れていること、トピックにおける生起確立 $p(w|z)$ が最大でも 0.1 付近であることがわかる。続いてトピック遷移構造の再構成機能により、モデムに関するトピックを表示させた。この際、図 5.6 の生起確立の推移を手がかりに、モデムに関連したトピックであると判断するしきい値を、トピックにおける「モデム」の生起確立が 0.01 以上であることと設定した。トピックの再構成結果を図 5.7 に示す。

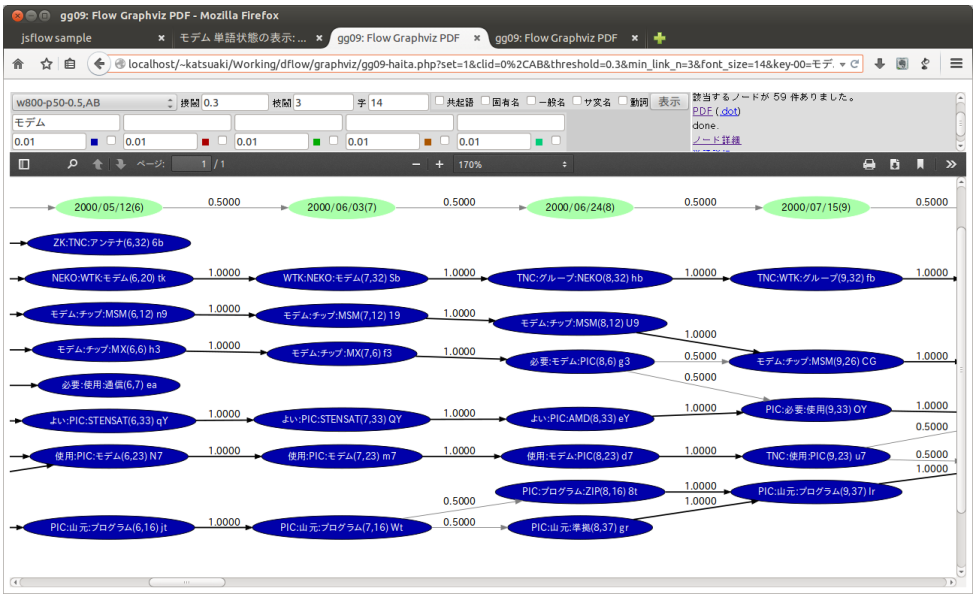


図 5.7 「モデム」に関連したトピック（一部）

次に、図 5.6、図 5.7 を元に、「モデム」がなぜ設計過程において文書集合中に現れなくなったかを調べる。

図 5.7 の表示と合わせて、4.4.4 に述べたトピック詳細の表示機能を用い、実際に議事録文書中のモデムに関する記述を確認した。この結果、モデムを実装するための「チップ」に関する議論、「PIC」を用いること、「TNC」をどのように設計するかの議論が行われていることがわかった。図 5.6 に「モデム」と共起する単語として、「チップ」「PIC」「TNC」が上位に現れていることとも合致する。

そこで、トピック遷移構造中での生起確率表示を用いてこれらの単語生起確率の推移を確認した。すると、「チップ」は「モデム」とほぼ同じ時期にしか使われていないこと（図 5.8）、「PIC」は「モデム」と重なる時期の出現確率が高いこと（図 5.9）、「TNC」は継続して現れること（図 5.10）が分かった。また、議事録中に「CARD-TNC（PC カードサイズのモデム）」など、モデムと TNC が等価のものである可能性を示唆する表現が見られたことから、「TNC」の意味を調べると [3]、Terminal Node Controller の略であり、モデムをその機能の一部として含むものであった。以上より、「TNC」の設計を検討することにより「モデム」に関する検討が終了したと推測できた。

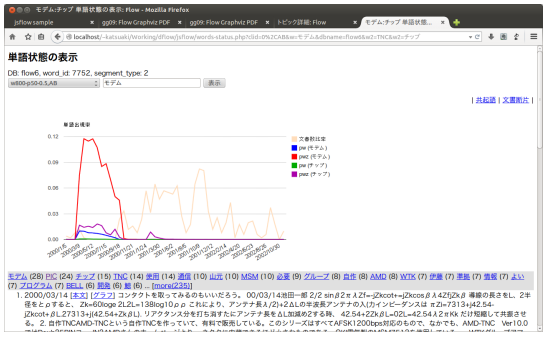


図 5.8 「モデム」「チップ」の生起確率推移

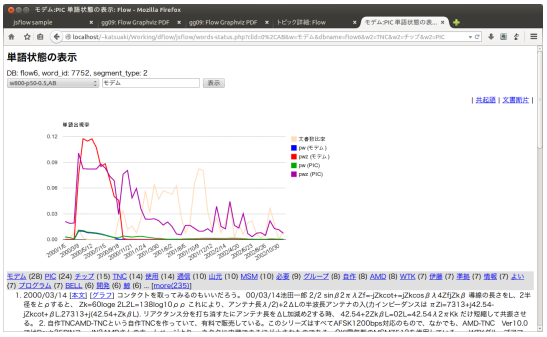


図 5.9 「モデム」「PIC」の生起確率推移

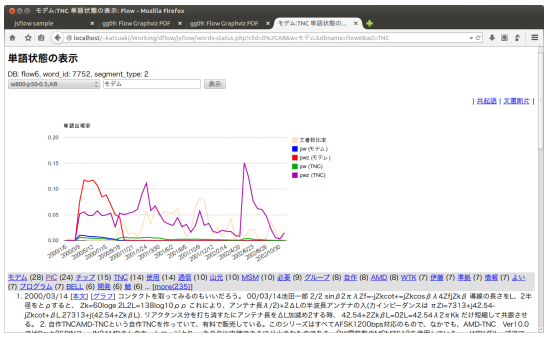


図 5.10 「モデム」「TNC」の生起確率推移

5.4.2　複数のキーワードによる再構成

CubeSat XI-IV の設計過程では、搭載無線機として、汎用品である TEKK-KS シリーズが設計初期に検討され、次に同じく汎用品の DJ シリーズが検討された。最終的に特注の無線機の開発に応じる西無線研究所に依頼し、専用設計の無線機を製作した。これらを確認するために、「TEKK」「DJ」「西無線」をキーワードとして指定しトピック遷移構造の再構成を実施した。力学的モデルによる表示システムを用いると、図 5.11 のような結果を得らた。この結果から、これらの無線機間の検討された順序関係、それらに付随して記述されている事項を概観することができた。また、トピックの詳細を確認することで、無線機に合わせて設計過程で検討された事項を確認することができた。

例えば、図 5.11 の表示、赤色のラベルのノードの分布から、西無線無線機の搭載が決定した後にも、DJ シリーズは文章中で引き続き言及されていることがわかる。各ノードを選択しトピックの詳細を表示、議事録詳細を参照すると、西無線無線機が完成するまでの間、代替として機能試験に DJ シリーズを用いたとの記述を見つけられた。

また、「DJ」と「西無線」両方に関連するトピックのうち、出現時刻が最も古いトピックに含まれる議事録を確認すると、西無線にどのような仕様で無線機の設計を依頼するかなど、西無線への無線機製作を依頼する際の検討事項の記述を見つけることができた。

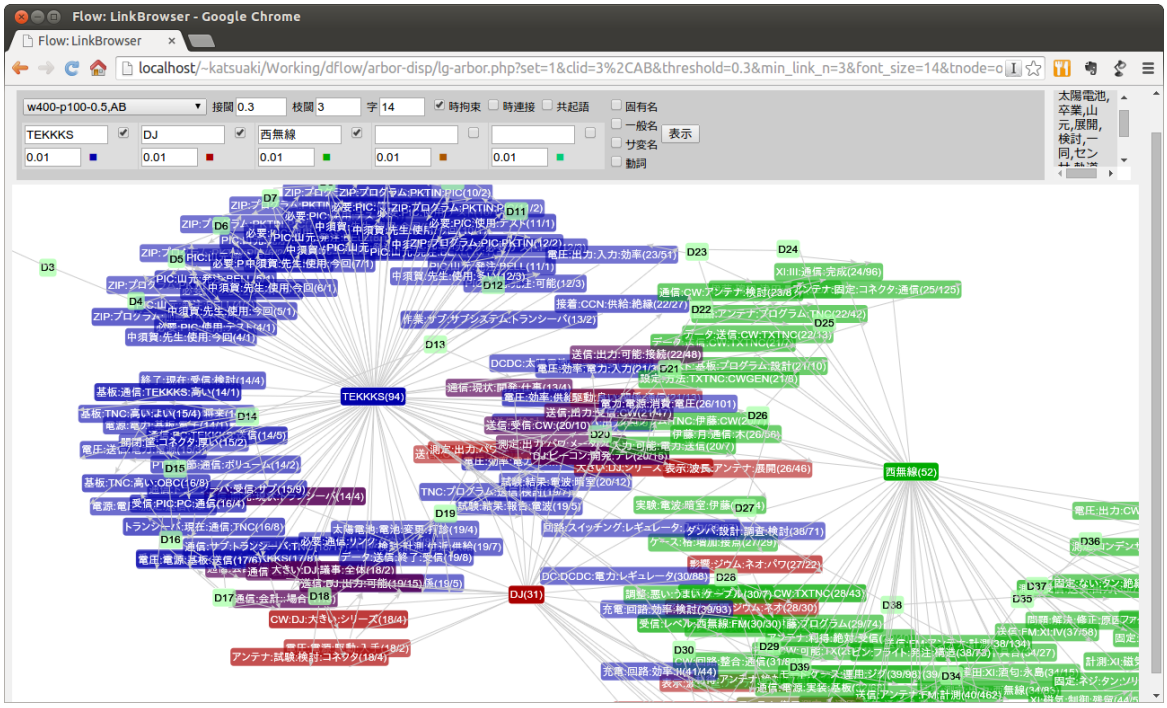


図 5.11 複数のキーワードによる再構成例（力学的モデルによる表示）

このように、単語によるトピック遷移構造の再構成・表示では、システム利用者が指定したキーワードに基づき、単数、あるいは複数のキーワードがどのような期間で出現したかを確認し、トピックの移り変わりを表示、確認することができることがわかった。

5.4.3　表示ノードラベルの指定

4.5.3 にて述べた、ラベル語を共起語とし品詞を指定する機能について確認を行う。例として、先に述べた無線機のひとつの「DJ」によりトピック遷移構造を再構成、ラベル語を共起語、かつ品詞としてサ変名詞を指定し得られた結果を図 5.12 に示す。

設定	設定値
時間方向の文書集合数	50
文書断片の長さ	400
トピック数	100
トピック忘却時の重み	0.5
トピック忘却時の新規文書判定手法	排他的クラスタリング

表 5.4 図 5.13 抽出用設定

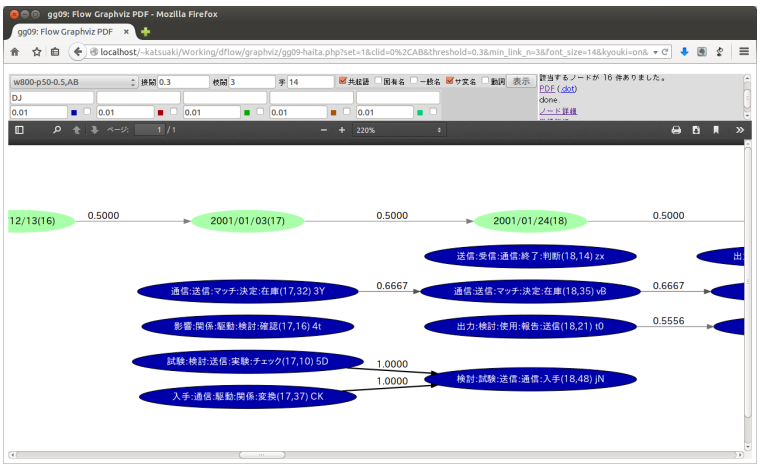


図 5.12 「DJ」と共起するサ変名詞（一部）・表 5.2 の設定による

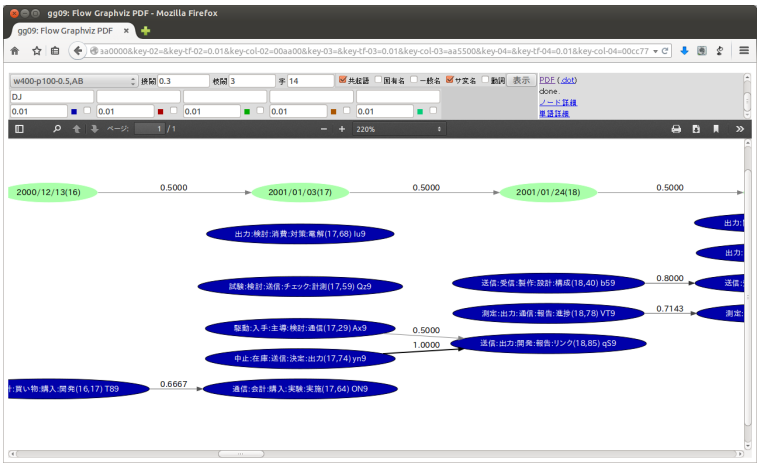


図 5.13 「DJ」と共起するサ変名詞（一部）・表 5.4 の設定による

単語	出現回数
送信	34
出力	23
検討	16
通信	15
測定	13
試験	11
受信	10
開発	10
報告	9
リンク	8

表 5.5 図 5.12・図 5.13 と同期間に「DJ」と共起するサ変名詞（上位 10 語）

キーワードとして文書集合に含まれる何らかの「着目対象」を指定すると、着目対象に関連するトピックをトピック遷移構造から抽出し、各表示手法により表示できる。また、抽出されたトピック遷移構造を表示をする際には、トピックを代表する単語をラベルとして表示する。本機能は、ラベルとして選択する単語の品詞を指定する機能である。例えば、品詞としてサ変名詞を指定すると、「〜する」となる名詞が抽出され、キーワードに対して行われた作業の概要が表示されることを期待できる。

たとえば、「DJ」に関連するトピックを抽出しラベル語をサ変名詞とした図 5.12 からは、「DJ」について「通信」「試験」「入手」などについて議論されていることがわかった。

さらに詳細を確認するために、表 5.4 のように断片長を短く（800bytes → 400bytes)、トピック数を増やし（50 → 100)、PLSI の結果を用いて排他的クラスタリングを行うことによりトピック遷移構造の抽出を行った後、同様に「DJ」に関連するトピックをサ変名詞をラベル語として抽出した例を、図 5.13 に示す。

図 5.12 と比較し図 5.13 では、「購入」「会計」「出力」「測定」などについて議論が行われていたことを確認できた。これは、文書断片の長さを短く、トピック数を多くしてトピック遷移構造の抽出を行うことにより、トピックの粒度が小さくなったこと、排他的クラスタリングを行うことにより、異なる作業を抽出できたことによると、考えられる。

図 5.13 の表示結果とあわせてトピックの詳細を確認することにより、無線機である DJ シリーズについて、「購入を検討していたこと」「出力に関して検討していたこと」「製造中止になり在庫の確認が必要であること」が並行して議論されていたことを確認できた。

表 5.5 に、図 5.13 と同じ期間に、「DJ」と共起するサ変名詞を、出現回数が多い順に上位 5 つを示す。表 5.5 からでもどのような作業が行われていたかはわかるが、それぞれの作業が行われたタイミングや、同じ作業を示すものなのかなどの関係はわからない。一方、トピック遷移構造を用いることにより、図 5.13 のように、同じタイミングで並行して行われている作業を分離して抽出することが可能である。

図 5.14 に、キーワードとして「DJ」を指定し、ラベル語の品詞として固有名詞を指定した例を示す。「DJ」と並行して「TEKKKS」が出現しており、先に述べたとおり、TEKK-KS シリーズと DJ シリーズとが並行して検討されていたことを読み取ることができる。

5.4.4　トピック遷移構造の抽出パラメータ設定

5.4.3 にて確認したように、トピック遷移構造を抽出する際のパラメータを変更することにより、その後の再構成で得られる結果も変化する。例えば、トピック数を変更することにより、再構成の対象とするトピックの粒度を変更することが可能である。

5.3 にて述べた俯瞰表示を行う際には、トピック数を少なめにし、トピックの粒度を大きくす

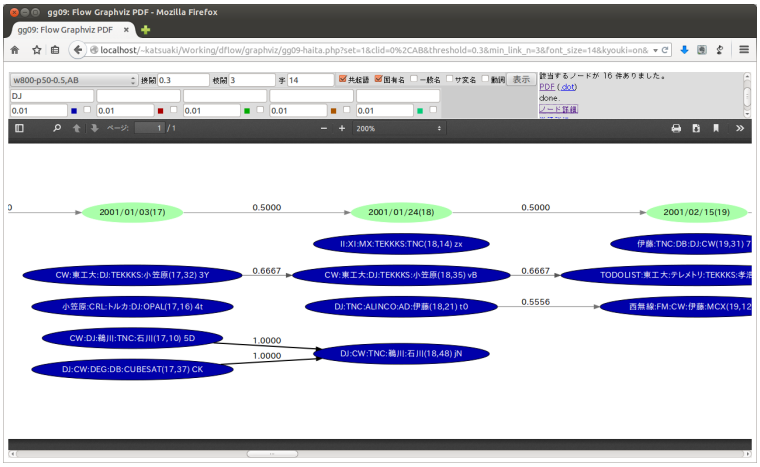


図 5.14 「DJ」と共起する固有名詞（一部）

るほうが全体の把握を行いやすかった。一方、5.4 にて述べた再構成と表示を行う場合、とくに、ラベル語の指定により、対象へ行われた作業を表示させる場合には、抽出するトピック数を多めにしてトピックの粒度を小さくすることにより、並行して議論されていた複数のトピックを確認することができた。

5.5　トピック遷移構造の提示システムの動作確認・分析のまとめ

本章では、4 章にて述べたトピック遷移構造の抽出・提示システムを小型人工衛星の設計プロジェクトの記録文書集合に適用し、各機能の確認を行った。確認を通じて、5.1 節に述べた、「人工衛星を構成する部分についての内容を文書集合から抽出して利用者に提示する」という目的を、システムが持つ複数の機能を通して果たせることを確認した。

まず、5.4.1 節に述べたキーワードによるトピック遷移構造の再構成と 5.4.3 節の表示ノードラベルの指定を組み合わせることにより、キーワードが指す対象に関する文書集合中の内容を、俯瞰的に確認することができた。この際、並行して行われた複数の記述を分離して確認することも可能であった。また、俯瞰的に確認した結果から、ある部品に関連すると思われるキーワード複数を拾い出すことができ、それらの生起確率の推移を確認することにより、例えば「モデム」が「TNC」へ置き換わったように、設計過程において検討された人工衛星を構成する要素の変遷を確認することができた。

さらに、5.4.2 節に述べた複数のキーワードによる再構成機能を用い、文書集合中でキーワードが重複する箇所の文書本文を確認することにより、要素間の変遷が起きた理由を見つけることができた。すなわち、「X という理由で A を B にかえた」の「A」「B」をキーワードとして再構成を行うことにより、「X」が記述されている可能性がある部分を拾い出すことが可能であった。

これらのキーワードの選択を行うために、トピック遷移構造の提示システムは、5.4.1 節の例

では、4.5.2 節で設けた単語入力時の提案表示、単語状態の表示などの機能を用い、関連する単語を選択する支援が有効であることを確認できた。また、トピック遷移構造中の単語生起確率の推移表示における共起語の表示も参考となった。単純な文書検索では、俯瞰的な表示機能を持たないため、関連するキーワードを拾い出す機能、キーワードの生起確率の推移を調べることはできず、このような設計過程の把握は難しいと考えられる。

これらの過程で、トピック遷移構造の抽出における文書の断片化長、抽出トピック数の設定を変更することによりトピックの粒度が変化すること、キーワードによる再構成を前提とする場合には、トピックの粒度をある程度小さくして、再構成結果を表示したほうが、文書集合中に記述されていることを具体的に把握しやすいなどの示唆を得た。

5.6　本章のまとめ

本章では、4 章で述べたトピック遷移構造の抽出と提示システムを実際の文書集合に適用し、システムの各機能の動作の確認と合わせ、得られる情報の分析を行った。各機能を組み合わせることにより、文書集合中に現れるさまざまな対象への記述の変化や、ある対象が別の対象へ入れ替わる様子などを把握することができた。

一方、「対象」の特定はシステムの利用者の興味を出発点としており、文書集合のどの部分が着目されるかは、システムの利用者により異なる。また、「部品が置き換わった」という現象はトピック遷移構造から把握することが出来るが、その「理由」は文書集合中の文書に記述されていないとわからない。すなわち、そもそも文書中に記されていないことは把握できない、という当然の限界が存在することも明らかになった。

第 6 章

さまざまな文書集合への適用

本章では、5 章で述べた小型人工衛星の設計議事録に続き、いくつかの時系列文書集合に対してトピック遷移構造の抽出・提示システムを適用する。対象とする文書集合とトピック遷移構造の抽出結果の概要について検討した後、各文書集合ごとに分析例を述べる。最後に、各文書集合の分析から得られた知見と 3 章の議論を踏まえ、文書集合ごとの特徴を議論する。

6.1　対象とする文書集合

トピック遷移構造の抽出は、文書集合がある程度以上の文書数からなること、文書が作成された時刻がわかることの 2 点を満たす文書集合を対象に行える。この 2 つの条件を満たす文書集合には、さまざま形態が考えられる。例えば、文書の作成に関わった人間の数に着目すると、会議の議事録のようにある程度同じメンバーが作成した場合、SNS の書き込みのように文書ごとに作成した人間がまったく異なる場合などが考えられる。

そこで、以下に述べるいくつかの蓄積形態や記述の対象が異なる文書集合を対象として、トピック遷移構造の抽出と、抽出したトピック遷移構造を用いた分析を行う。

6.1.1　審議会議事録 1: 中央教育審議会議事録

議事録例の 1 つとして、文部科学省中央教育審議会 [68] の第 1 回（2001 年 2 月 1 日）〜第 124 回（2020 年 1 月 24 日）までの議事録を取り上げる。参加した各委員ほか出席者の発言を記録した議事録である。

本文書集合は教育という抽象的な事項に関する会議の記録である。中央審議会の下部組織としていくつかの分科会、さらにその下には部会・少委員会が置かれており、具体的な議論はこれらの下部の会議で行われているものと考えられる。

6.1.2　審議会議事録 2: 中央環境審議会地球環境部会議事録

中央教育審議会の議事録に対し、もう少し細かい議論を行っている議事録の例として、環境省に置かれている審議会の中の部会のひとつ、中央環境審議会地球環境部会の議事録を取り上げる。対象とするのは第 1 回（2001 年 2 月 16 日）〜第 145 回（2020 年 8 月 4 日）までの議事録である。

地球環境部会の所掌事務は「地球環境の保全に係る重要な事項に関すること。」[52] とされており、議論の内容も、対象である地球環境に対してどのような行為をするか直接的な決定を行い結果を確認することが目的ではなく、地球環境に対してどのように臨むかの議論を行うことが目的であると考えられる。議事録を確認すると、この会議は地球環境に関して広い範囲に触れており、メンバー以外にも外部の有識者を招いて意見を聴くなど、記述されている内容も多岐に渡る。

6.1.3　ツイート集合：2014 年前半の「人工知能」検索結果

次の例として、Twitter から特定の語を含むツイートを検索し、得られたツイートからなる文書集合を取り上げる。

ここでは、2013 年 12 月から 2014 年 6 月までに Twitter から単語「人工知能」を含む公開ツイートの検索結果を用いる。人工知能学会会誌 2014 年 1 月号表紙の描写内容に関する言及が多く含まれることを期待して Twitter からツイートを検索、収集できたツイートを文書集合とした [56]。

6.2　中央教育審議会議事録

6.2.1　トピック遷移構造の抽出

中央教育審議会議事録を対象に、トピック遷移構造の抽出を行い、得られた特徴語、および得られたトピックの 2 次元配置表示を行い、おおよその抽出結果を確認する。

前処理

中央教育審議会の Web サイト [68]、および国立国会図書館にアーカイブされた中央教育審議会の Web サイト [4] から、第 1 回（平成 13 年 2 月 1 日）から第 124 回（令和 2 年 1 月 24 日）の HTML 形式で記述された議事録のファイルを取得した。取得した HTML ファイルからテキストを抽出後、1 つの発言を 1 つの文書とした。この結果、7,310 個の文書が作成され、異なり単語数は 12,185 語、平均文書長は 1,164bytes であった。なお、文書の作成日には会議の開催日を設定した。

抽出結果概要

表 6.3 に表 6.1 の設定により、表 6.4 には表 6.2 の設定により、トピック遷移構造の抽出を行った結果から、各 $C_{n,i}$ について $p(C_{n,i})$ が大きい順に上位 10 個のトピックについて、$p(w_k|C_{n,i})$ が最大となる単語 w_k、すなわち各時間区間において出現確率が上位 10 のトピックにおける、最も特徴的な単語を示す。

また、4.4.5 節で述べた手法によりアニメーション表示させたうち、全体、$n = 10, 20, 30, 40, 50$ の状態を、図 6.1 に表 6.1 の設定で抽出した場合、図 6.2 に表 6.2 の設定で抽出した場合をそれぞれ示す。

図 6.1、図 6.2 の 2 次元にレイアウトしたトピックの位置を確認すると、小型人工衛星設計議事録（図 5.2）と異なり、どちらの図も時間の経過に伴った変動を観察しづらい。

一方、図 6.1、図 6.2 を比較すると、図 6.1 は、一部散らばっているトピックがあるものの、それ以外は領域にばらつかず固まっている。表 6.1 に示した特徴語も、ほぼ同じ単語が並んでいることから、トピック間の類似性が、表 6.1 の設定の方が表 6.2 の設定よりも高いと考えられる。

図 6.2 を細かく確認すると、$n = 10$ の左側中央に見られた「教養：人間：社会」「言語：文化：理解」が時間を経るとともに表示されなくなっている。一方、中央右側の「地域：社会：連携」「学校：環境：整備」付近のトピックは、$n = 10$ では少ないが $n = 9$ などでは確認できる。とはいえ、トピックの出現傾向が、図 6.2 の表示上、右側に移動している様子は伺える。

設定項目	設定値
時間方向の文書集合数	50
文書断片の長さ	800bytes
トピック数	50
トピック忘却時の重み	0.3
トピック忘却時の新規文書判定手法	確率合計・0.5 以下

表 6.1 表 6.3、図 6.1 の抽出用設定

設定項目	設定値
時間方向の文書集合数	50
文書断片の長さ	400bytes
トピック数	50
トピック忘却時の重み	0.3
トピック忘却時の新規文書判定手法	確率合計・0.5 以下

表 6.2 表 6.4、図 6.2 の抽出用設定

D_n	特徴語									
2	教育	分科	問題	教育	活動	資料	教育	教養	日本	意味
4	教育	審議	教育	教育	基本	人	意見	学校	科学	活動
6	審議	教育	学校	問題	基本	社会	大学院	教育	資料	社会
8	基本	社会	大学院	教育	教育	教育	審議	今	教育	教養
10	教育	教育	教育	教育	教育	教育	教育	教育	教育	教育
12	教育	教育	教育	教育	教育	教育	教育	教育	教育	教育
14	教育	教育	教育	教育	教育	教育	教育	教育	学校	学校
16	教育	教育	教育	教育	教育	基本	教育	教育	学校	教育
18	教育	教育	教育	教育	学校	教育	教育	教育	教育	教育
20	教育	教育	教育	教育	教育	教育	教育	教育	教育	教育
22	教育	教育	教育	教育	教育	教育	教育	教育	教育	教育
24	審議	教育	教育	教育	社会	大学	地方	子ども	議論	学習
26	教育	大学	基本	意見	教育	学校	教育	教員	教養	指導
28	教育	教育	審議	大学	教育	教員	教育	教育	教育	大学院
30	教育	審議	安全	大学	教育	教員	学習	学校	資料	教育
32	教育	審議	大学	教員	支援	社会	教育	活動	人	先生
34	教育	教育	今	教員	教育	社会	大学	大学	教育	学校
36	教育	教育	資料	教育	教育	教育	関係	教育	部会	基本
38	社会	審議	地域	教育	教育	教育	教員	教育	教育	教育
40	教育	教育	教育	教育	教育	教育	教育	教育	教育	教育
42	教育	教育	教育	教育	教育	教育	教育	教育	教育	教育
44	教育	教育	学校	教育	教育	教育	教育	教育	教育	教育
46	教育	教育	教育	教育	教育	教育	教育	教育	教育	教育
48	教育	教育	教育	教育	教育	教育	教育	教育	教育	教育
50	教育	学校	審議	教育	大学	教育	学習	教育	大臣	活用

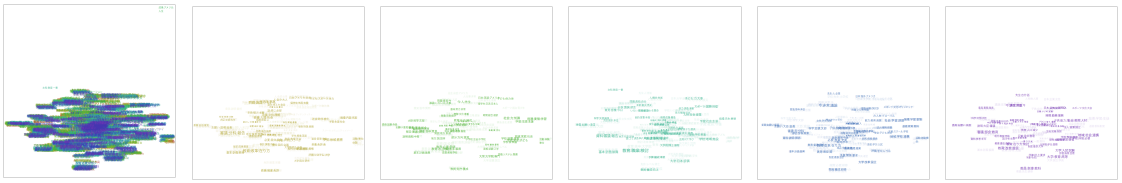
表 6.3 中央教育審議会議事録の上位 10 トピック特徴語（n = 偶数のみ、設定は表 6.1）

D_n	特徴語									
2	教育	分科	活動	教員	検討	教養	社会	問題	大学	審議
4	教育	問題	社会	教員	学校	今	大学	学校	審議	分科
6	審議	教育	問題	基本	分科	部分	教員	検討	学習	学校
8	教育	資料	教育	教育	教育	問題	学校	基本	大学院	指導
10	審議	教育	問題	学校	教養	基本	必要	今	言葉	教員
12	問題	審議	基本	学校	教育	課題	教育	人間	大学院	大学
14	教育	問題	学校	必要	人	教育	お願い	部分	人間	基本
16	審議	検討	問題	地域	社会	教育	基本	意味	義務	教員
18	教育	審議	問題	教養	基本	今	検討	大学院	学校	教員
20	教育	今	教育	社会	指導	報告	地域	会長	教育	大学
22	教育	お願い	教育	大学	教育	学校	大学	教育	議論	計画
24	教育	審議	学校	基本	問題	検討	教育	説明	地方	大学
26	教育	審議	大学	必要	学校	教員	基本	委員	大臣	子供
28	審議	大学	教育	教育	議論	お願い	教員	教育	大臣	計画
30	教育	資料	課題	学校	大学	基本	人材	スポーツ	学習	教員
32	審議	学習	教員	教育	学校	大学	基本	教育	教育	委員
34	教育	審議	基本	教育	支援	今	社会	教育	教員	具体
36	教育	資料	教育	教育	教育	教育	今	道徳	大事	大学
38	審議	教育	教育	今後	教育	非常	意見	能力	教育	予算
40	今	教育	審議	必要	地域	社会	科学	お願い	教員	予算
42	教育	教育	審議	大臣	社会	在り方	学校	大学	問題	教育
44	教育	教育	審議	学校	教員	社会	資料	大学	指導	問題
46	教育	教育	教育	教育	教育	教育	教育	教育	教育	教育
48	教育	審議	地域	社会	指導	教育	検討	非常	議論	社会
50	審議	大学	教員	地域	教育	社会	教育	大学	今	説明

表 6.4 中央教育審議会議事録の上位 10 トピック特徴語（n = 偶数のみ、設定は表 6.2）

全体　$n = 10$　$n = 20$　$n = 30$　$n = 40$　$n = 50$

図 6.1 中央教育審議会議事録のトピック 2 次元配置（設定は表 6.1）



全体　$n = 10$　$n = 20$　$n = 30$　$n = 40$　$n = 50$

図 6.2 中央教育審議会議事録のトピック 2 次元配置（設定は表 6.2）

6.2.2　分析

表 6.2 の設定で抽出したトピック遷移集合について、4 章で述べたトピック遷移構造の提示システムを用いて、含まれる情報の分析を行った。

表 6.3、表 6.4 には「教育」「学校」などの単語が多く、文書集合の変化の概要をこの表から行うことは難しい。そこで、いくつかの内容に焦点を宛て分析を行う。まず、分析対象の 1 つ目として、どちらの表にも定期的に出現している「大学院」を対象にすることにする。

続いて、分析する候補を見つけるために、もう少し詳細な内容を確認する。そのために、表 6.2 に示した設定で抽出したトピック遷移構造に対し、4.4.1 節で述べた静的なグラフ表示機能を用い、トピック間のリンクのしきい値を 0.2 以上として表示を行わせた。図 6.3 に、得られた表示のうち、文書集合開始時期の一部を示す。

図 6.3 では、「教育」「検討」などの他、「免許」「制度」「学習」「生涯」に関するトピックの存在がわかる。そこで、「教員免許更新制度」と「生涯学習」の 2 つを「大学院」に加えた、3 つを内容を分析対象とする。

分析例：大学院

はじめに、トピック遷移構造における単語「大学院」の生起確率の推移を図 6.4 に示す。この図より、「大学院」が文書集合の全期間に渡って出現していることがわかる。

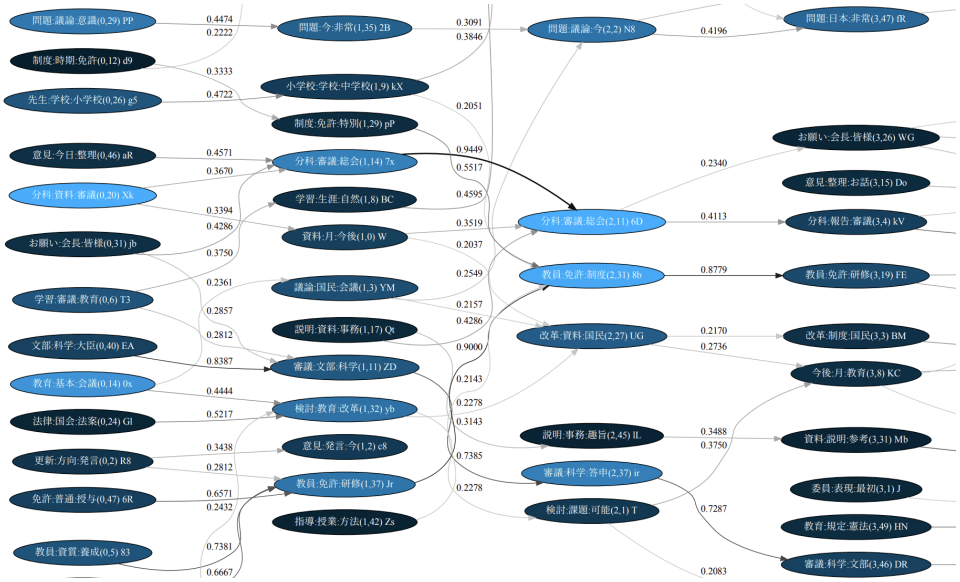


図 6.3 中央教育審議会議事録のトピック遷移構造の静的グラフ表示（一部）

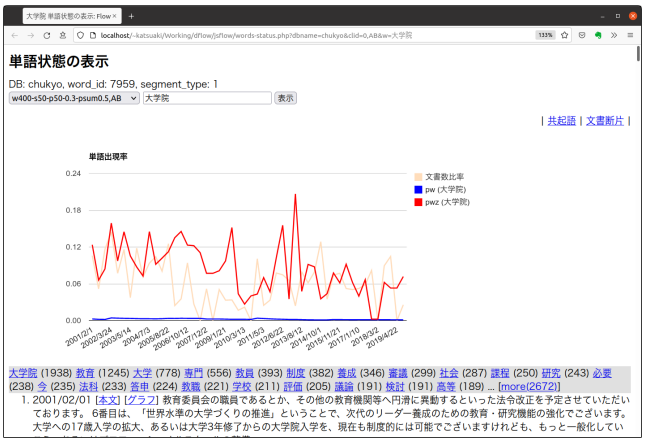


図 6.4 「大学院」の生起確率推移

次に、「大学院」を含むトピックの連なりにどのようなものがあるかを確認するため、連接する類似トピック間のリンクのしきい値を 0.2 として、4.4.3 節に述べた力学的モデルにより表示を行わせた（図 6.5）。なお、トピックとキーワードの関連を判定するしきい値には、特別な記述を行わない限り以下では 0.01 を用いる。この結果、何らかの答申に関連したトピックの連なり、研修・経営に関連したトピックの連なり、教育プログラムに関連したトピックの連なりなどを確認できた。

続いて、「大学院」という単語のみで、関連するトピックのみを再構成し、サ変名詞をラベルとして表示させ、記述の推移を確認する（図 6.6）。図左端のトピックを確認すると、上から順に、「海外への頭脳流出防止のための魅力改善」「人材高度化と大学院卒業者の失業」「教育の質の保証と大学・大学院接続」「教職大学院による教員の質の向上」「教員養成と大学・大学院の関係」

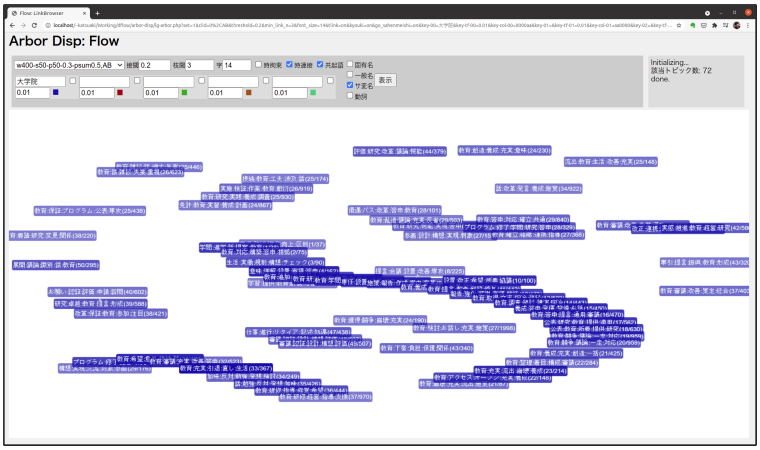


図 6.5 「大学院」に関連するトピック

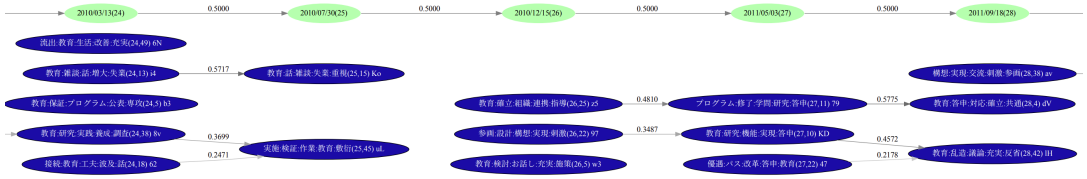


図 6.6 「大学院」に関連するトピック（一部）

が記述されていた。またこれらの記述は、「教職生活の全体を通じた教員の資質能力の総合的な向上方策についての諮問」と「大学分科会」からの報告を受けてのものであった。

また、図を確認すると、「プログラム」が左端（2010 年 3 月）と右から 2 番め（2011 年 5 月）の両方に出現している。議事録本文を確認すると、どちらも同一の発言の一部であった。しかし、2010 年 3 月のトピックは「大学教育の学位プログラム」、2011 年 5 月のトピックは「学位プログラムとしての一貫した博士課程教育」について述べており大学院との連携に関する記述、大学院の教育内容に関する記述と、それぞれ内容が異なると考えられる。

左から 3 つ目（2010 年 12 月）の 3 つのトピックでは、「グローバル人材・教職人材の育成」「グローバル・イノベーション人材の育成」「キャリア教育」について述べられ、関連する右端（2011 年 9 月）のトピックでは、「人材育成のために法科大学院を設けたが乱造され過ぎでは」といった意見の記述を確認できた。

分析例：教員免許更新

教員免許更新制度は、2009 年度から開始され、2021 年 8 月に更新制度廃止の方針が中央教育審議会で議論されている。まず、「教員」「免許」「更新」をキーワードとして 4.5.1 節に述べたトピック遷移構造の再構成を実施した結果を、図 6.7 に示す。

図 6.7 に含まれるトピックを確認したところ、「更新」を含まず「教員」「免許」のみを含むト

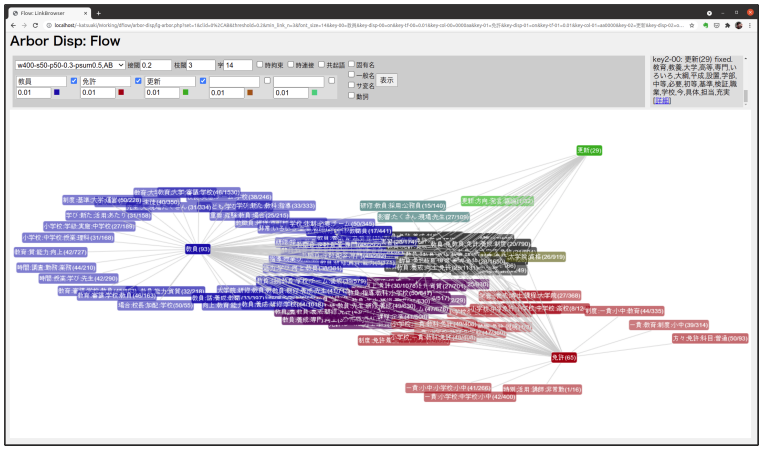


図 6.7 「教員」（青）「免許」（赤）「更新」（緑）による再構成結果

ピック（青と緑のみからリンクされるトピック）は、ほぼ見られなかったため、以後の確認では、「免許」「更新」のみをキーワードとする。

次に、「免許」「更新」を含むトピックを 4.4.2 節に述べた時間軸を固定した表示を用いて共起するサ変名詞をラベルとして表示させた結果の、2007 年 7 月〜2011 年 9 月部分を、図 6.8 に示す。

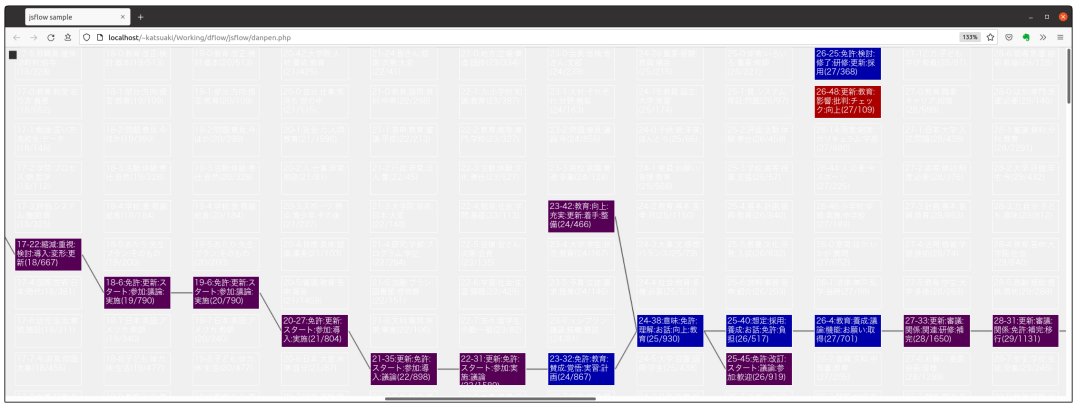


図 6.8 「免許」（青）「更新」（赤）を含むトピックの時間軸固定表示（2007 年 7 月〜2011 年 9 月）

図 6.8 のトピックの詳細を順に確認したところ、教員免許更新制度の「導入」「スタート」の報告、すでに更新講習に「トライ」している大学が存在することの報告を確認できた。「議論」「参加」は、「議論へ参加した立場から……」との発言が複数行われており、新たな議論の喚起を示す内容ではなかった。

なお、「更新」を含まないトピックでも、「教員の質の向上」に意見が表明されており、その後に続く 2011 年 1 月には、「（教員の向き・不向きを）更新制があれば途中でチェックできる」といった発言も記録されていた。この発言に対し、同じ会議中に、「（免許更新制度が）本当に実質化されているか」という発言が別の委員から行われていた。また、これらの発言が含まれている議事録全体を確認したところ、これらの発言は、「教員の資質能力向上特別部会の審議状況」の

報告に対して行われていることを確認できた。

続いて、表示する期間を処理対象とした議事録の末尾、2020 年 1 月を含む部分とした表示を図 6.9 に示す。なお、「更新」が含まれるトピックと判断する生起確率のしきい値を 0.01 とした場合、関連するトピックの表示がなかったことから、しきい値は 0.001 に変更した。

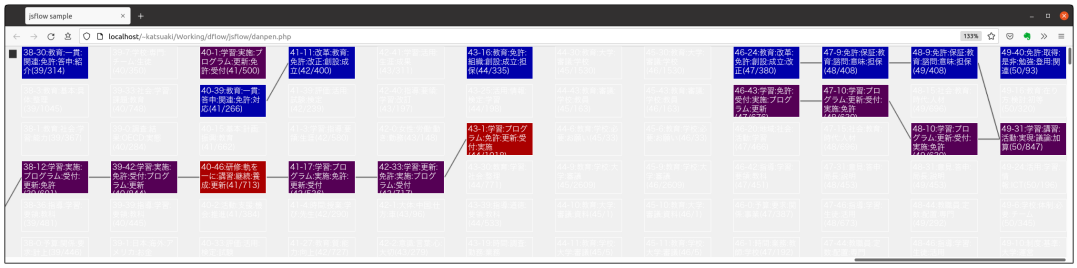


図 6.9 「免許」（青）「更新」（赤）を含むトピックの時間軸固定表示（2015 年 7 月〜2020 年 1 月）

「学習」「実施」トピックには、教員免許更新講習を実施実績があることの報告があるのみであったが、右端の「学習」「加算」を含むトピックには、「学校における働き方改革推進本部にて（中略）現在の仕組みを、10 年間の教師の学びを蓄積し、加算する方式に抜本的に転換」とする発言が含まれていた。

分析例：生涯学習

図 6.10 に、「生涯」「学習」をキーワードとしてトピック遷移構造を再構成した結果を示す。この図より、「生涯」に関連するトピックは、ほぼ全て「学習」にも関連していることがわかる。よって、以後、「生涯」のみをキーワードとして「生涯学習」についての分析を行う。

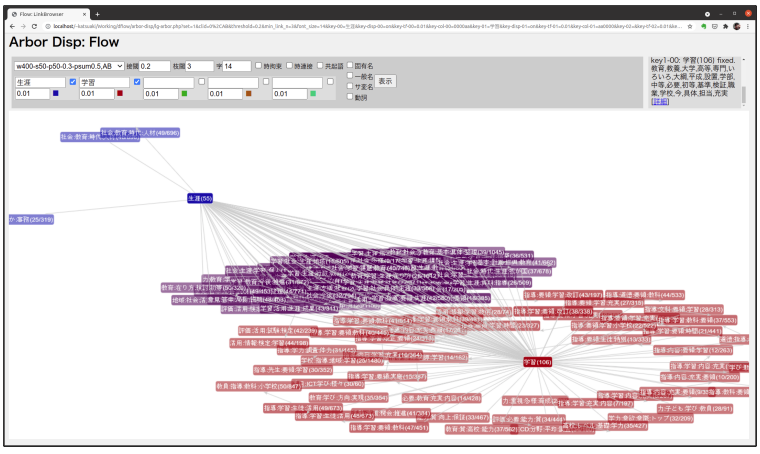


図 6.10 「生涯」（青）「学習」（赤）による再構成結果

「生涯」を含むトピックを 4.4.2 節に述べた時間軸を固定した表示を用いて共起するサ変名詞をラベルとして表示させた結果のうち、2006 年 10 月〜2010 年 12 月の部分を、図 6.8 に示す。

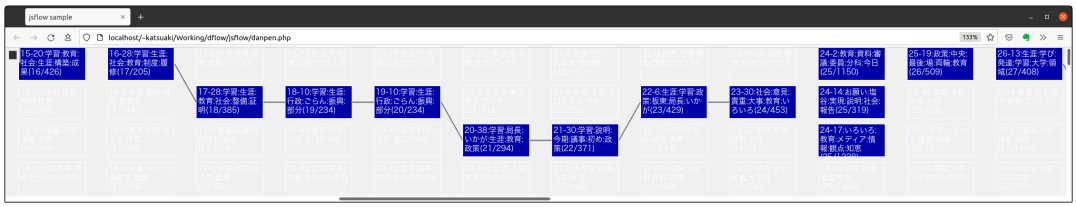


図 6.11 「生涯」を含むトピックの時間軸固定表示（2006 年 10 月〜2010 年 12 月）

図 6.12 「生涯」（青）「答申」（赤）「諮問」（緑）による再構成結果

表示されたトピックを確認していくと、ほとんどのトピックが、生涯学習分科会での審議をまとめた報告、諮問内容の説明、答申内容の説明のいずれかであった。

そこで、生涯学習学に関するトピックと「答申」「諮問」とのトピックの重複を確認するために、「生涯」「答申」「諮問」を指定して 4.5.1 節に述べたトピック遷移構造の再構成を実施した（図 6.12）。

図 6.12 から、「答申」と「諮問」には両者にまたがるトピックが存在するが、「生涯」と「答申」にまたがるトピックはほとんど存在しないことがわかった。「答申」に関するトピックを確認したところ、「〜についてご意見をいただきながら議論を進めてきました」といった、「答申」を行うことそのものについてのトピックが抽出されており、「答申内容」には触れていなかった。トピック遷移構造の抽出設定を表 6.2 から表 6.1 に変更し、断片の長さを 2 倍にしても、同様の結果だった。

なお、図 6.11 にて同時にトピックが 3 つ出現している部分は、2010 年 1 月 21 日の第 71 回審議会にて、自由な意見交換が 1 時間ほど行われた結果、「生涯学習に関する制度の説明」「委員への呼びかけ（お願いします、ほかにいかがでしょうか）」以外に、「生涯学習」と「男女共同参画」や「メディア教育」の関わりに関する意見が述べられていた。

6.2.3 考察

中央教育審議会は、文部科学大臣による「諮問」として示された内容について、「答申」としてこれに回答するために、下部の分科会を含めて議論を行い、中央教育審議会にて最終答申を確認する、という流れで進められている。各回の会議の議事録は、文部科学省の担当部局や審議会の下部に位置づけられる分科会で作成された資料を確認し、審議会に参加している委員が意見を述べる、という形で構成されていた。審議会が扱う範囲は教育行政全般に渡るが、各回で取り扱う範囲は、事前に準備された資料によりおおよそ限定されていた。

審議会内部で議論した結果が、次の開催時に反映されるわけではないため、議事録の記述内容は、会議内での議論の進展に沿って会を進めるごとに順を追って変化するのではなく、各回で用意された資料に基づき、その説明内容と委員の表明する意見が変化する、という形であった。

また、説明の導入、委員の意見表明に対する導入、など議事進行が丁寧に行われるために、説明の導入部分だけで一定以上の文字数があり、例えば 6.2.2 節に述べたように、「答申」に関するトピックは、答申内容の説明の前ふりのみからなる、という事象が確認された。

トピック遷移構造の抽出では、文書集合中の記述対象は、複数が継続的に出現することを前提とし、処理対象の文書集合中に言及がないトピックを古いトピックとする処理を行う。また、類似するトピックかの判定は、時間的に隣接しているトピック同士でしか行わない。

一方、中央審議会の議事録では、文書集合内には言及がないが、会議外部で継続して議論が行われ、その結果が議事録に現れる、という形であった。そのため、期間をおいて類似した内容が出現すると考えられるが、トピック遷移構造ではこの類似性を拾うことができなかった。

今後、中央教育審議会に関して、本研究手法により議論の変化を詳細に追うためには、中央教育審議会議事録のみではなく、諮問・答申 [67] や、下部の分科会・部会の議事録も含めた文書集合を作成し、文書集合内部に議論の進展を含むように準備を行うことが考えられる。

6.3 地球環境部会議事録

6.3.1 トピック遷移構造の抽出

地球環境部会議事録を対象に、トピック遷移構造の抽出を行い、得られた特徴語、および得られたトピックの 2 次元配置表示を行い、おおよその抽出結果を確認する。

前処理

環境省中央環境審議会地球環境部会の Web サイト [51] から、2001 年 2 月から 2020 年 8 月までの HTML 形式で記述された議事録のファイルを取得した。取得した HTML ファイルからテ

キストを抽出後、1 つの発言を 1 つの文書とした。この結果、8,287 個の文書が作成され、異なり単語数は 14,423 語、平均文書長は 1,381bytes であった。なお、文書の作成日には会議の開催日を設定した。

抽出結果概要

表 6.7 に、表 6.5 の設定によりトピック遷移構造の抽出を行った結果から、各 $C_{n,i}$ について $p(C_{n,i})$ が大きい順に上位 10 個のトピックについて、$p(w_k|C_{n,i})$ が最大となる単語 w_k、すなわち各時間区間において出現確率が大きいトピックにおける、特徴的な単語を示す。

4.4.5 節で述べた手法によりアニメーション表示させたうち、全体、$n = 10, 20, 30, 40, 50$ の状態を図 6.13 に表 6.5 の設定で抽出した場合、図 6.14 に表 6.6 の設定で抽出した場合について、それぞれ示す。

設定項目	設定値
時間方向の文書集合数	50
文書断片の長さ	800bytes
トピック数	50
トピック忘却時の重み	0.3
トピック忘却時の新規文書判定手法	確率合計・0.5 以下

表 6.5 表 6.7、図 6.13 の抽出用設定

設定項目	設定値
時間方向の文書集合数	50
文書断片の長さ	400bytes
トピック数	50
トピック忘却時の重み	0.3
トピック忘却時の新規文書判定手法	確率合計・0.5 以下

表 6.6 表 6.8、図 6.14 の抽出用設定

D_n	特徴語									
2	合意	資料	議定	環境	制度	削減	委員	対策	途上	日本
4	対策	検討	議定	環境	途上	削減	アメリカ	準備	取組	委員
6	問題	委員	見直し	合意	制度	資料	委員	議定	削減	日本
8	海洋	対策	排出	部会	議論	資料	必要	日本	技術	質問
10	対策	目標	排出	今	制度	非常	途上	資料	意見	議論
12	対策	対策	対策	対策	対策	対策	対策	委員	委員	委員
14	委員	対策	対策	委員	対策	対策	対策	対策	委員	対策
16	対策	対策	対策	対策	対策	対策	対策	対策	委員	委員
18	議論	対策	対策	対策	委員	対策	対策	対策	議論	議論
20	対策	対策	委員	対策	対策	委員	今	対策	委員	委員
22	議論	委員	対策	問題	対策	環境	必要	エネルギ	委員	委員
24	環境	必要	委員	対策	対策	議論	対策	対策	対策	対策
26	排出	対策	対策	議論	議論	議論	委員	排出	議論	議論
28	エネルギ	対策	議論	対策	対策	対策	排出	環境	議論	対策
30	議論	エネルギ	エネルギ	対策	議論	議論	議論	議論	議論	対策
32	議論	エネルギ	対策	議論	議論	委員	議論	対策	エネルギ	議論
34	エネルギ	議論	議論	委員	エネルギ	議論	エネルギ	議論	議論	議論
36	対策	議論	対策	議論	委員	エネルギ	エネルギ	エネルギ	対策	対策
38	委員	対策	対策	委員	対策	計画	委員	委員	委員	対策
40	委員	計画	対策	委員	計画	環境	計画	対策	対策	対策
42	対策	対策	計画	対策	委員	対策	計画	委員	資料	目標
44	対策	技術	適応	資料	パリ	対策	議論	エネルギ	目標	地域
46	エネルギ	対策	適応	評価	フロン	エネルギ	目標	パリ	委員	今後
48	イノベーション	適応	部門	計画	目標	指摘	戦略	環境	エネルギ	カーボン
50	指摘	適応	社会	戦略	評価	削減	気候	議論	日本	指摘

表 6.7 地球環境部会議事録の上位 10 トピック特徴語（n = 偶数のみ、設定は表 6.5）

D_n	特徴語									
2	議定	環境	検討	意見	削減	制度	委員	今	排出	対策
4	アメリカ	制度	環境	削減	会合	取組	委員	検討	対策	日本
6	問題	削減	議定	アメリカ	検討	目標	部門	地球	委員	対策
8	部門	対策	資料	非常	途上	削減	報告	海洋	委員	国際
10	対策	評価	今	日本	資料	部門	意見	議定	排出	委員
12	対策	排出	目標	お願い	議論	部門	審議	国際	非常	施策
14	議論	計画	必要	非常	日本	対策	産業	今	施策	税
16	対策	問題	部分	国際	議論	税	温暖	資料	お願い	委員
18	対策	議論	問題	必要	排出	国際	目標	部会	電力	日本
20	対策	社会	議論	非常	意見	問題	自動車	目標	計画	排出
22	必要	議論	エネルギ	対策	排出	資料	検討	部門	計画	委員
24	基本	検討	問題	目標	自動車	社会	制度	環境	省エネ	エネルギ
26	議論	基本	日本	排出	環境	京都	意見	対策	問題	エネルギ
28	議論	排出	基本	必要	対策	今	部会	技術	意見	エネルギ
30	部分	環境	モデル	議論	部会	エネルギ	制度	話	京都	委員
32	検討	日本	議論	排出	必要	議論	エネルギ	部会	コスト	政策
34	対策	意見	今	モデル	問題	環境	社会	委員	制度	目標
36	社会	目標	対策	今	環境	エネルギ	コスト	適応	議論	国際
38	議論	日本	環境	社会	削減	対策	エネルギ	部分	委員	約束
40	非常	削減	計画	対策	策定	適応	エネルギ	国民	COP	目標
42	排出	国際	エネルギ	目標	点検	環境	対策	計画	月	非常
44	適応	対策	削減	エネルギ	気候	委員	目標	排出	資料	COP
46	排出	適応	適応	気候	長期	エネルギ	委員	目標	方向	プライシング
48	議論	戦略	適応	委員	エネルギ	気候	イノベーション	目標	国際	カーボン
50	対策	資料	議論	社会	適応	非常	報告	削減	戦略	委員

表 6.8 地球環境部会議事録の上位 10 トピック特徴語（n = 偶数のみ、設定は表 6.6）

全体　$n = 10$　$n = 20$　$n = 30$　$n = 40$　$n = 50$

図 6.13 地球環境部会議事録のトピック 2 次元配置（設定は表 6.5）

全体　$n = 10$　$n = 20$　$n = 30$　$n = 40$　$n = 50$

図 6.14 地球環境部会議事録のトピック 2 次元配置（設定は表 6.6）

図 6.13 では、$n = 10$、$n = 50$ 以外では、トピックが中央に固まってレイアウトされている。表 6.7 から、トピックの特徴語もほぼ同じ単語が並んでいることを確認できる。

図 6.14 では、若干の変化はあるものの、全期間に渡ってトピックがほぼ散らばっており、全体的な大きな変化は無いことを確認できる。表 6.8 からも、横方向に同じタイミングで出現している単語の種類を確認すると表 6.7 に比べて多く内容が散らばって見えるが、縦方向に全期間に渡ってみると、特徴語に大きな変化は無いことがわかる。

6.3.2　分析

続いて、抽出したトピック遷移集合について、4 章で述べたトピック遷移構造の提示システムを用いて、含まれる情報の分析を行った。

分析例 1：温室効果ガス排出削減の評価

地球温暖化防止のために、温室効果ガスの排出削減を目指し、削減量などの評価が行われているものと考え、その様子を議事録から調べることとした。

「温室」「効果」「ガス」「排出」「削減」の 5 つのトピックの重なり、4.4.3 節に述べた力学的モデルによりを確認した（図 6.15）。その結果、「温室」「効果」「ガス」と「削減」のみまたがるトピックは非常に少なく、「排出」と「削減」にほぼ含まれることがわかった。そこで、「排出」と「削減」、「評価」をキーワードとして、以後の分析を進めることとした。

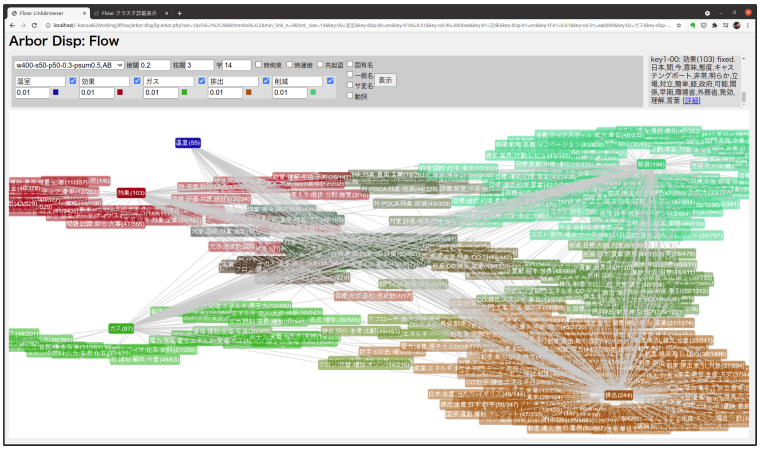


図 6.15 「温室」「効果」「ガス」「排出」「削減」による再構成結果

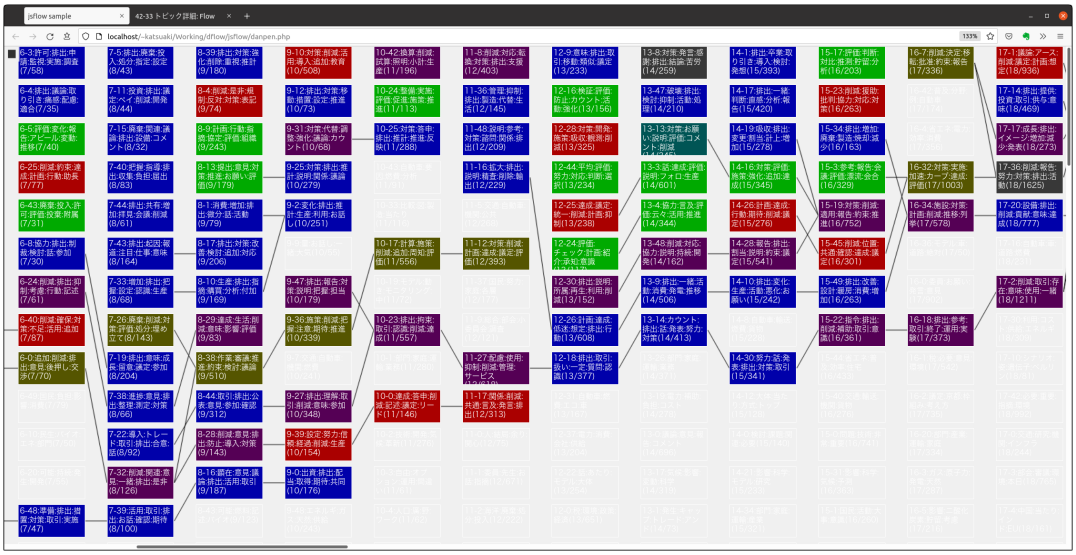


図 6.16 「排出」（青）「削減」（赤）「評価」（緑）を含むトピックの時間軸固定表示（2003 年 6 月〜2007 年 9 月）

続いて、「排出」「削減」「評価」を含むトピックを 4.4.2 節に述べた時間軸を固定した表示により確認した。ラベルを共起するサ変名詞として確認を行った画面の例を図 6.16 に示す。「排出」「削減」「評価」が重なるのは図 6.16 で灰緑色に表示されているトピック、例えば「6-0 追加：削減:排出:…」である。

図 6.16 の中ほど、2005 年ごろまでは、「地球温暖化大綱」に基づく評価に関する記述があるが、それ以後は、評価に関する記述は連続せずに、間隔をあけて出現する。評価の枠組みの変化を期待して確認を進めていくと、2011 年 10 月には「『日本の技術による削減貢献量に対して応分の評価を付与する仕組』の構築 」に関する記述が、2018 年 2 月には「経済産業省関係の 37 の対策・施策に紐付く 114 の対策評価指標」の記述があった。

また、「評価」のみに関連するトピックを確認したところ、2020 年 2 月の議事録から「気候変動影響評価報告書」をまとめていること、「気候変動適応法」が制定されていることがわかった。そこで、気候変動適応法についてあらためて調査することとした。

分析例 2：気候変動適応法

分析例 1 を受けて、気候変動適応法に関連する議事録中の記述を確認する。「気候変動適応法」に関連するトピックを確認するため、キーワードとして入力しようとしたところ、4.5.2 節に述べたキーワードサジェスト機能により、図 6.17 のように「気候」のみが表示され、「気候変動適応法」「気候変動」は 1 つの単語としては認識されていないことが確認できた。

図 6.17 「気候変動適応法」の入力過程における単語のサジェスト表示

そこで、「気候」「変動」「適応」「法」と 1 単語ずつをキーワードとし、4.4.1 節の静的なグラフ表示を行わせ、議事録全体における関連するトピックを表示させた（図 6.18）。また、図 6.18 下段の「気候」「変動」「適応」と関連するトピックが連続している部分について、ラベル語をサ変名詞として時間軸固定表示させた結果を、図 6.19 に示す。

これらに表示されたトピックを元に議事録を参照していくと、「気候変動枠組条約」の議論が第 1 回議事録から行われており、国際的な合意に至るための検討の報告が行われていた。その後、国内での「計画」「方針」などの議論が増え、2014 年の第 122 回では世界各国の国内での気候変動適応への取り組みや関連法の成立状況が紹介されいてる。2015 年には「政府全体としての適応計画を策定」することが述べられており、2018 年の議事録には、「(適応計画にあわせて) 平成 32 年ごろを目途に第 2 次気候変動影響評価を実施」することなどが述べられていた。

一方、「気候変動適応法」については、2018 年第 139 回議事録にて国会での成立と 12 月からの施行に伴い実施される事項の説明がなされているのみであった。

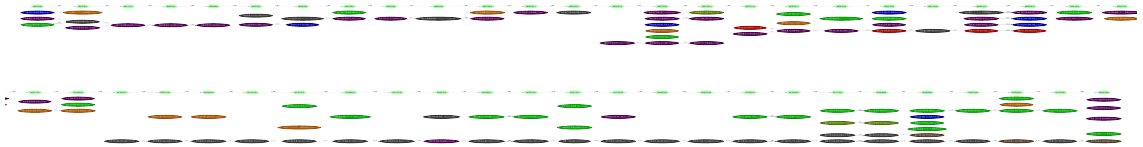


図 6.18 「気候」(青)「変動」(赤)「適応」(緑)「法」(橙) に関連するトピックの静的な表示 (上段から下段へ続く)

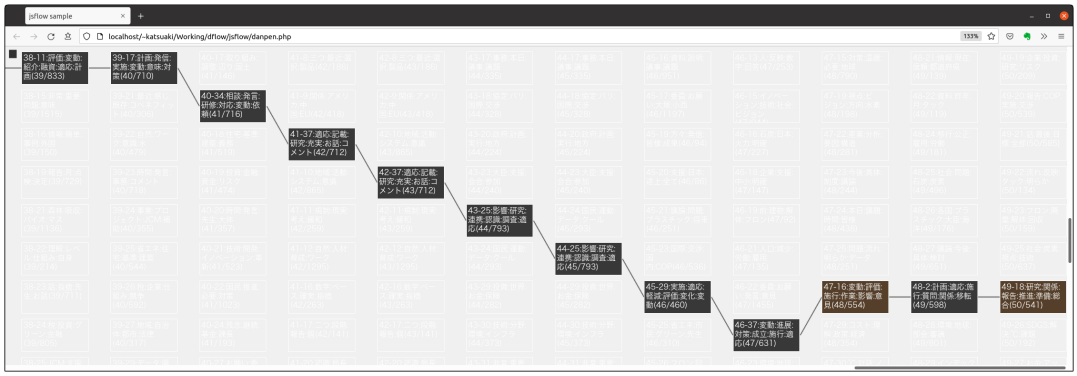


図 6.19 「気候」(青)「変動」(赤)「適応」(緑)「法」(橙) に関連するトピックの時間軸固定表示 (2015 年 12 月〜2020 年 8 月)

6.3.3 考察

地球環境部会は、中央環境審議会が持つ 8 つの部会の 1 つであり、下部に 7 つの小委員会を持つ (2021 年 9 月現在)。所掌事務は「地球環境の保全に係る重要な事項に関すること。」とされており、主に地球温暖化に関する議論が議事録には記されている。

6.2 節で取り上げた中央教育審議会に対し、審議会の中の 1 つの部会であるためか、扱う領域が限られ、より具体的な記述が見られた。しかし、5 章で扱った小型人工衛星設計議事録のように部品などのレベルに対応する用語が定義されるわけではないため、議論の対象が 1 つの単語で表わされず、トピックの再構成のキーワードとして直接指定が行えないなどの問題が明らかになった。

トピック遷移構造の抽出前に、1 つの単語として形態素解析器の辞書に登録すれば、1 単語として処理することは可能である。しかし、どこまで事前に文書集合を精査し、単語登録を行うかの調整が新たな課題となる。

6.4　ツイート集合：2014 年前半の「人工知能」検索結果

人工知能学会誌 2014 年 1 月号以降の表紙をめぐる Twiitter 上の記述を扱う。同号以降の人工知能学会誌の表紙は、女性蔑視ではないかと Twitter 上で炎上した [60]。文字列「人工知能」を Twitter 上から検索、得られたツイートを対象の文書集合とした。この「炎上」をツイート発言者の関係により分析する研究がなされている [66]。

6.4.1　トピック遷移構造の抽出

ツイート収集と前処理

Twitter の Search API を用いて「人工知能」を検索キーにし、ツイートを検索、収集した。これにより、2013 年 12 月 25 日 19 時付近から 2014 年 6 月 6 日 18 時付近（どちらも日本時間）までの 235,979 件のツイートを収集した。続いて、公式リツイートを除外した 131,522 件から、処理量を減らすために約 $\frac{1}{3}$ にあたる 43,862 件をランダムに選択した。また、前処理として、「RT」「QT」を含む以後の文字列と URL を除去した。ツイートの長さは 140 文字が上限であることから各ツイートの断片化は実施せず、ツイート全体をひとつの断片に相当するものとして、トピック遷移構造の抽出処理を行った。

抽出結果概要

表 6.9 に、表 6.10 の設定でトピック遷移構造の抽出を行ない得られた結果のうち、各時間区間において出現確率が上位 10 のトピックにおける最も特徴的な単語を示す。また、得られたトピック遷移構造を、4.4.5 節で述べた手法によりアニメーション表示させたうち、全体、$n = 10, 20, 30, 40, 50$ の状態を図 6.20 に示す。

表 6.9 から、人工知能学会の表紙に関する議論が主である（「人工知能学会」「差別」「批判」「女性」など）ことがわかる。他に、「GOOGLE」が複数現れている。文書集合から Google を含むツイートを検索したところ、Google が DeepMind を買収したなどのニュース記事を示すツイートが複数あることを確認できた。

図 6.20 の下部から中央にかけての領域に、「女性」「批判」「表紙」などを特徴語に含むトピックの分布を確認できた。上部の領域は「GOOGLE」「(IBM の）ワトソン」など人工知能関係のニュースが、中央上部右よりには、人工知能を活用したロボット、投資システムなどの半ば広告のようなトピックの分布が確認できた。時間の経過を通してみると、これら広告のようなツイートの割合が徐々に増加していると考えられる。

D_n	特徴語									
2	人工知能	批判	人工知能学会	女性	男	掃除	イメージ	まとめ	差別	人
4	ロボット	ロボット	学会	人工知能学会	女性	人工知能	掃除	差別	問題	人間
6	ロボット	まとめ	学会	問題	人工知能学会	女性	人工知能	人間	掃除	差別
8	男	人工知能	批判	問題	人工知能学会	女性	差別	人間	蔑視	ニュース
10	批判	学会	人工知能	問題	まとめ	差別	人工知能学会	話	掃除	人工知能
12	男	表紙	人工知能	批判	問題	学会	掃除	性	表紙	人
14	ロボット	批判	人工知能	差別	人工知能学会	学会	女性	GOOGLE	ロボット	掃除
16	ロボット	差別	人工知能	まとめ	女性	批判	人工知能学会	男性	GOOGLE	人間
18	男	男	人工知能	問題	学会	差別	GOOGLE	人	絵	人工知能
20	男	差別	人工知能	男	女性	批判	男性	人	研究	学会
22	女性	女性	人工知能	女性	女性	学会	人間	人工知能学会	今	話
24	批判	女性	批判	学会	人工知能	差別	掃除	女性	人間	人
26	人工知能学会	人工知能学会	学会	女性	人工知能	件	人間	表紙	人工知能	人工知能
28	男	人工知能学会	人工知能	学会	差別	女性	掃除	問題	GOOGLE	人工知能
30	男	男	学会	人工知能学会	人工知能	差別	人工知能	GOOGLE	人工知能	女性
32	女性	女性	女性	人工知能	女性	差別	学会	GOOGLE	ニュース	人
34	女性	人工知能	学会	男	人工知能学会	表紙	人工知能	問題	人工知能	ルンバ
36	批判	人工知能	男	人工知能学会	差別	学会	メイド	批判	ルンバ	人間
38	男	人工知能	女性	人工知能学会	人工知能	学会	人間	ルンバ	GOOGLE	脳
40	批判	人工知能	人工知能	学会	問題	女性	ルンバ	差別	人間	人工知能
42	男	人工知能	学会	人工知能	差別	人間	人工知能学会	ルンバ	人工知能	人工知能
44	まとめ	人工知能	学会	表紙	男	人間	ルンバ	GOOGLE	人工知能	人工知能
46	男	女性	人工知能学会	人工知能	人工知能	人工知能	人工知能	学会	人間	表紙
48	女性	人工知能	ルンバ	女性	問題	学会	人間	SFP	人工知能	人
50	ロボット	人工知能	問題	SFP	女性	学会	ルンバ	GOOGLE	人間	研究

表 6.9 「人工知能」を含むツイートの上位 10 トピック特徴語（n = 偶数のみ、設定は表 6.10）

設定項目	設定値
時間方向の文書集合数	50
文書断片の長さ	ツイート全体
トピック数	50
トピック忘却時の重み	0.3
トピック忘却時の新規文書判定手法	確率合計・0.5 以下

表 6.10 表 6.9、図 6.20 の抽出用設定

全体　$n = 10$　$n = 20$　$n = 30$　$n = 40$　$n = 50$

図 6.20 「人工知能」を含むツイートのトピック 2 次元配置（設定は表 6.10）

6.4.2　分析

抽出したトピック遷移集合について、4 章で述べたトピック遷移構造の提示システムを用いて、含まれる情報の分析を行った。

分析例：ジェンダー・蔑視

表 6.9 にトピックの特徴語の一覧を示したが、もうすこし詳細に内容を確認するために、4.4.1 節で述べた静的なグラフ表示機能を用い、トピック間のリンクのしきい値を 0.2 以上としてグラフを描画させた。得られたグラフのうち、ツイート収集開始時期の一部を図 6.21 に示す。

図 6.21 から、「人工知能学会（誌）の表紙と女性」「デザイン変更」「女性蔑視」「ジェンダー」（形態素解析の影響で「ジェン」と「ダー」に分離されている）などのツイートの存在が想定される。そこで、関係していそうな「ジェンダー（ジェン）」と「蔑視」をキーワードとして 4.5.1 節に述べたトピック遷移構造の再構成を実施した結果を、図 6.22 に示す。

図 6.22 より、ジェンダーと女性蔑視に関するツイートには重複がほぼ無いことがわかる。図でジェンダー・蔑視のトピック集合をつないでいるのは、時間的に隣接し類似度が高いトピックである。これら 2 つのトピックを確認したところ、「人工知能表紙問題を含むジェンダー勉強会の開催」、「人工知能学会誌 2014 年 3 月号での 1 月号を振り返る特集」に関するツイートが含まれており、両者を接続する記述を確認できた。

なお、トピック遷移構造の抽出・再構成に問題がないかを確認するため、「ジェンダー」「蔑視」

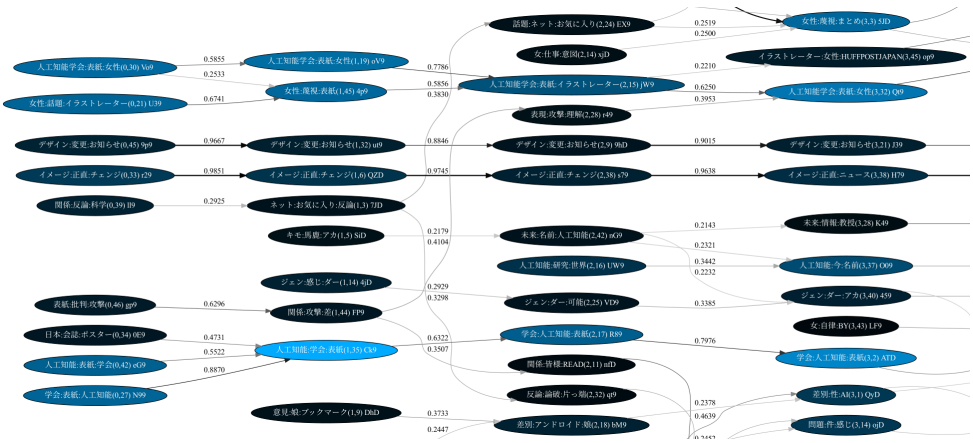


図 6.21 「人工知能」を含むツイートのトピック遷移構造の静的グラフ表示（一部）

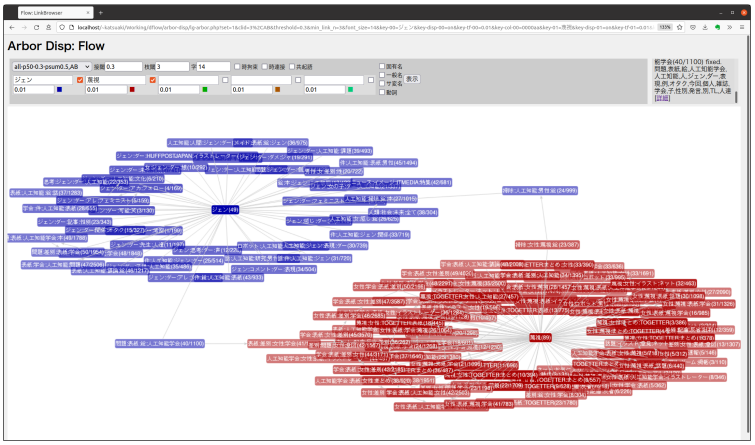


図 6.22 「ジェンダー（ジェン）」（青）「蔑視」（赤）による再構成結果

から類推される「差別」も含めて再構成を実施した（図 6.23）。この結果、「蔑視」と「差別」が重複するトピックの存在を確認でき、トピック遷移構造の抽出・再構成処理に問題はないと考えられる。

「蔑視」に関連するトピックに含まれるツイートを確認すると、「蔑視」をタイトルに含む外部のまとめ記事、ニュースサイトなどへのリンクを紹介しているツイートがほとんどであった。同じタイミングに複数の「蔑視」を含むトピックが出現しているのは、紹介先が異なることにより紹介に含まれるタイトルが異なることに起因していた。

「ジェンダー」に関連するトピックに含まれるツイートには、人工知能学会の表紙がジェンダーに絡む問題であることを述べたさまざまなツイートが含まれており、特定の記事を紹介するなど、出来事に触れたツイートはほとんど見られなかった。同じタイミングで複数の立場の意見が表明されていることから、異なるトピックに分類されることが望ましい。しかし、図 6.24 に示したとおり、「ジェンダー」を含むトピックは同じタイミングは 1 つしか出現しておらず、トピックの分離が行われていない。これは、「ジェンダー」を含むツイートが文書集合中で 192 件

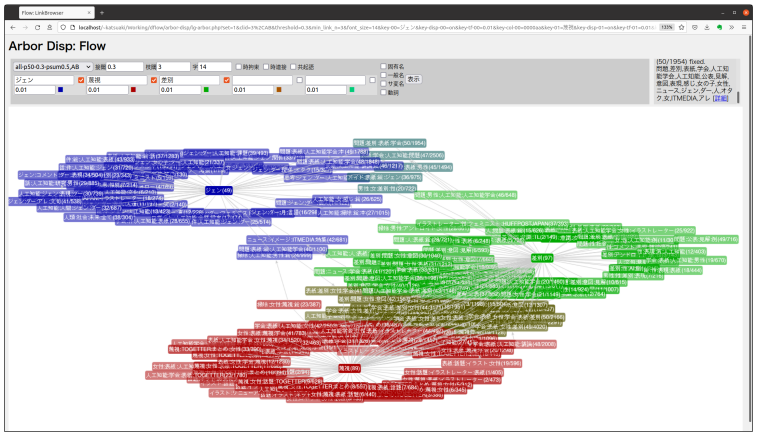


図 6.23 「ジェンダー（ジェン）」（青）「蔑視」（赤）「差別」（緑）による再構成結果

と少なく（「蔑視」を含むツイートは 1,238 件）、他のツイートと比較して複数のトピックへ分類されるほどの数がなかったこと、ツイートは 140 文字以下と短いためにジェンダーに対する立場をトピック抽出では分離できなかったことなどが、原因として考えられる。

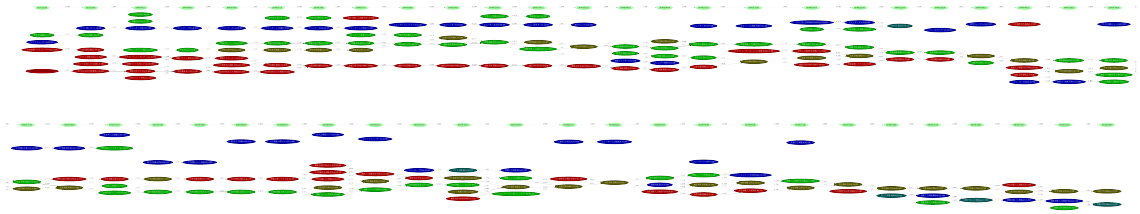


図 6.24 「ジェンダー（ジェン）」（青）「蔑視」（赤）「差別」（緑）による再構成結果の静的な表示（上段から下段へ続く）

さらに、「ジェンダー」「別紙」「差別」の 3 つの単語による再構成結果を 4.4.1 節で述べた静的なグラフ表示により表示した結果を、図 6.24 に示す。初期に「蔑視」に関するトピックが複数存在するが、それ以後は、各単語を含むトピックが継続的にほぼ均等に出現しており、トピックの中身を確認したところ、時間経過に沿った大きな変化は見られなかった。

分析例：ロボット

分析例として、表 6.9 にも見られた「ロボット」に関する記述の推移を確認する。図 6.25 に、「ロボット」を含むトピックを 4.4.2 節に述べた時間軸を固定した表示を用いて表示させた一部を示す。

図 6.25 では、トピックの特徴語として、「表紙」「女性」「批判」を含むものが多く見られる。これらのトピックの詳細を確認すると、例を図 6.26 に示すように、トピック中の生起確率が上位の 1〜2 つのツイートがそれぞれ異なる内容で、それ以外のツイートはまとめ記事へのリンク

図 6.25 「ロボット」を含むトピックの時間軸固定表示（一部）

図 6.26 「ロボット」を含む「女性」「批判」「表紙」を特徴語とするトピックの例（一部処理）

を示すものであった。

その他、「アシモフ」を含むトピックを確認すると、人工知能学会誌 2014 年 3 月号の表紙や小特集に関するニュース記事に言及したツイート、関連してアシモフの小説を想起したことを述べたツイートを確認できた。「無料」「動画」を含むトピックでは同内容の広告ツイート、「過去」「生命」を含むトピックは「人工知能と人工生命」という書籍を紹介する同内容のツイートが、多数含まれていた。

このように、「ロボット」を含むツイートの内容を時間の経過に沿って確認できるが、それぞれが対象としている「ロボット」は、異なるものであった。

6.4.3　考察

ツイート集合から抽出したトピック遷移構造では、出来事を反映したツイートがほとんどであった。

対象としたツイート集合からは、まとめサイトやニュースサイトなどに掲載された「記事」に対するツイートが多数確認できた。また、これらのツイートの記述内容は記事を紹介するだけであり、全期間に渡って大きな変化はなかった。一度つくられた「記事」はその時点での出来事を述べたものであり、それ以後変更されない。また、「記事」の作成は Tiwtter の外部で行われ、「記事」の作成に関する議論はツイート集合中には含まれない。これらの理由により、一度「記事」が作られると、ツイートの内容の変化がそこで止まり、同内容のツイートが繰り返されていると考えられる。

また、異なるタイミングで同一の単語を特徴語とするトピックでも、同一の対象についての記述ではなく、同じ単語で表現される異なる対象についての記述が行われている事例を確認できた。

時間経過に沿って変化する対象への記述を見つけられなかった理由として、対象としたツイート集合は、Twitter の Search API を用いてキーワード検索により集めたものであることが考えられる。ツイートは 140 文字という上限があるため、同一の内容について、同じ単語を含ませず続けてツイートする、あるいはリプライを返すことが多い。このように、命題を提示した後に続けて議論がなされていた場合には、キーワードを含まないツイートに議論部分が含まれ、キーワードを含むツイートを収集した今回の手法では集めることができない。このため、時間の経過に沿って変化する対象の抽出には、検索結果の前後のツイート、リプライ関係にあるツイートなども含めて収集し、トピック遷移構造の抽出対象とすることが考えられる。

6.5　文書集合ごとの性質の検討

本節では、5 章で述べた小型人工衛星設計議事録、本章で述べた審議会議事録、ツイート集合へ、4 章で述べたトピック遷移構造の抽出・再構成システムの適用を通した結果と、3 章の議論を踏まえつつ、扱った文書集合の性質について、議論を行う。

小型人工衛星設計議事録からは、時間の経過を通して、設計対象とする部品の入れ替わりや、その理由などを確認することができた。一方、審議会議事録やツイート集合からは、特定の内容に関する記述の変化を確認することはできたが、他の記述対象との移り変わりや、その理由などを確認することができなかった。

そこで、各文書集合について、文書の蓄積形態と文書に記述される対象の検討を行う。

6.5.1 小型人工衛星設計会議議事録

5 章で扱った小型人工衛星設計の記録は、複数のメンバーが、小型人工衛星を対象とし、衛星を打ち上げて宇宙空間での写真撮影を行うという目的のために行われた設計過程を記録した文書である。文書の作成・蓄積期間全体を通して具体的な対象が存在し、対象への行為を行う人間が 1 人以上、記録を行う人間も 1 人以上存在する。また、設計に関して作成された文書はほぼ全て、処理対象とした文書集合に含まれている。

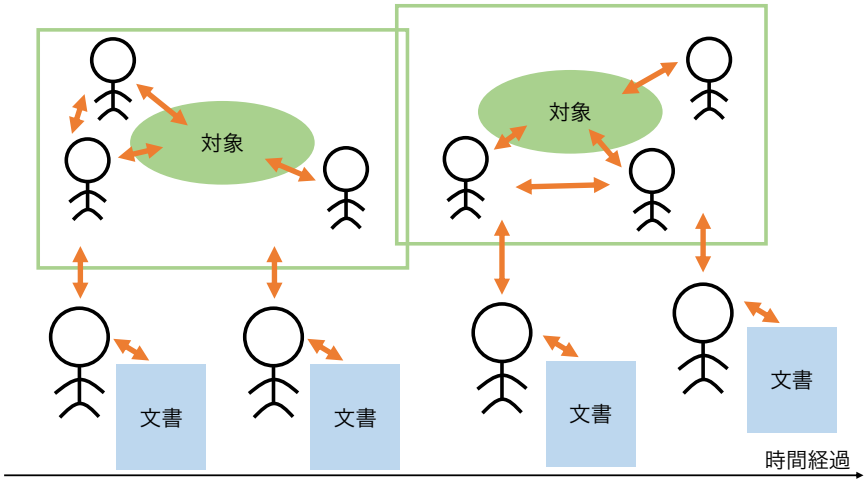


図 6.27 設計会議議事録の蓄積形態

図 6.27 に設計会議議事録における、対象と人間、時間経過の関係を示す。設計過程では、設計を行う人間の行為の結果、設計の対象は時間経過に沿って変化させられ、その変化結果を受けて人間が新たな行為を行う。設計会議の議事録は、設計の時間経過に沿って、対象の状態や対象への行為、それらについての人間の検討内容を記録したものとなる。

6.5.2 審議会議事録

中央教育審議会議事録、地球環境部会議事録の両審議会の議事録は、複数の人間がある内容についての議論を繰り返した記録である。会議において発言を行う人間が複数人、記録を行う人間も 1 人以上存在する。

図 6.28 に、会議議事録型の文書群における、対象と人間、時間経過の関係を示す。

議事録の元となる会議は、参加メンバーが対象を確認する場であり、対象への行為が発言という形でなされている。会議の対象は、実在する具体的なものであることもあれば、「会議に参加するメンバーが共有しようとしている認識」などの仮想的な存在であることもある。

会議の目的が対象について議論を行うことであるため、対象は、会議と別に存在するのではな

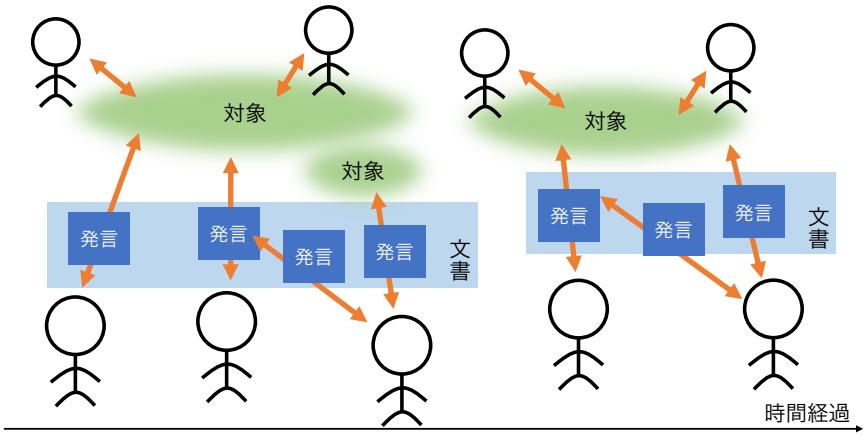


図 6.28 審議会議事録の蓄積形態

く、会議の中に存在すると考えられる。そのため、会議を何回か行う間に実時間が経過しても、対象は影響を受けない可能性がある。

また、会議にて行われる対象についての議論は、対象に関する情報の共有のみを行う場合、議論過程で対象へ新たな変更を加えようとする場合など、さまざまな場合が考えられる。前者の場合、会議の外部で対象に変化が発生しない限り、会議を繰り返しても対象に関する文書中の記述には変化はない。後者の場合でも、変更を加えた結果の議論がなければ対象の記述には変化が発生せず、文書集合中の記述には時間の経過に沿った変化は現れない。

このように、審議会の議事録では、実在の対象を外部に持たない可能性、対象に変更を加えない可能性があり、このような文書集合の蓄積過程では、類似した内容が繰り返し記録される可能性が高いと考えられる。

6.5.3　ツイート集合

Twitter から特定のキーワードを検索し集めたツイート集合は、多数の人間が、それぞれ自身の考える対象について、独自の記述を行った文書の集合である。図 6.29 に、ツイート集合における、対象と人間、時間経過の関係を示す。

一部のツイートは、キーワードに関連してほぼ同一の対象を対象を仮想的に共有している可能性はある。また、ツイートへの返信ツイートとして議論が続く場合もあるが、返信ツイートがキーワードを含まず、検索により集めたツイート集合に含まれない可能性が高い。また、単純なツイート数の割合としても、返信ツイートよりも独立したツイートのほうが数が多い。

すなわち、ほとんどのツイートは記述した人間が関心を持つ対象について述べたものであり、複数のツイート間で、時間経過に沿って対象が共有されることは、ほぼ無いと考えられる。

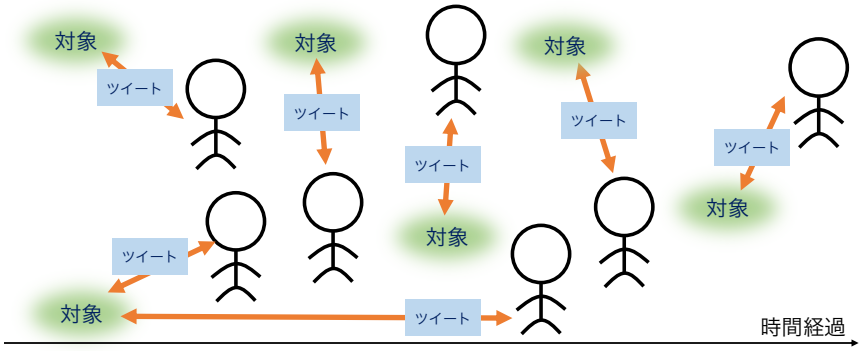


図 6.29 ツイート集合の蓄積形態

6.5.4　文書集合の性質の検討のまとめ

審議会議事録では、文書の記述対象は、事前に存在する具体的な事物・事象以外に、会議の場で共通認識としようとする仮想的なものの場合がある。そのため、記述対象が複数回の会議では異なるものとなる。また、具体的に存在する対象の場合が記述されている場合も、対象への操作や変化は審議会の外で起きており、文書には変化結果を確認する形の記述が行われていた。

ツイート集合では、主たる記述対象は外部で発生した事象であり、その発生自体を述べた文書（ツイート）が多く、また、共通する対象への意見などの記述を行っていても、他の文書との関連性がほぼ存在しなかった。

一方、小型人工衛星設計議事録では、設計対象の人工衛星を構成するさまざまな要素・要素の集合が具体的な対象として記されており、複数の文書を通して継続的に記述対象とされていた。

以上のことより、時間の経過を通して記述の変化を確認するためには、複数の文書にまたがり共通して存在し、その変化が記述されている対象が必要である、と考えられる。

6.6　本章のまとめ

本章では、時間の経過に沿って蓄積された文書集合、審議会の議事録、Twitter 検索結果といった蓄積のされ方が異なる文書集合を対象とし、トピック遷移構造の抽出を行い、複数の被験者による提示システムの利用結果を確認した。

2.2 節に挙げた既存のトピックの遷移抽出や抽出結果の提示手法は、抽出対象とした文書集合全体におけるトピック遷移の位置づけの俯瞰提示を行う。また、対象として、ニュース記事や学術論文を扱っており、結果を見る人間が処理対象の文書集合の内容に対しある程度の知識を持つことが、暗黙の前提であると考えられる。

例えば、6.4 節で取り上げた「人工知能」を含むツイートでは、「人工知能学会学会誌の表紙の

掃除をする女性型アンドロイドについて議論が起きたこと」を知っていれば、俯瞰的な表示中の「家事」「批判」などの単語の表記で何が議論されているかわかるが、知らなければ「人工知能」と「家事」「批判」という単語だけから、これらの関係を類推することは難しい。

一方、トピック遷移構造の提示システムを用いると、複数のトピックが時系列に沿って提示され、トピックの概要から詳細をたどり、トピック含まれるツイートをシステムを操作しながらひとつずつ確認し、記述内容の移り変わりを確認することが可能である。このように、システムの利用者が興味を持ったトピックやその特徴語から関連する文書の詳細を読み進めることにより、内容に関する事前知識がなくても、理解可能な部分から、文書集合を読み始めることができた。

本章ではまた、利用結果を踏まえて、対象とした文書集合の性質の検討の行った。これにより、「複数の文書にまたがり変化する対象が存在」しない場合、文書集合の記述内容からは時間の経過を読み取ることが難しいことを示した。

第 7 章

通時的対象の抽出と利用

本章では、時系列に沿って蓄積される文書集合の中で、変化の中心となる「通時的対象」を定義し、文書集合中から通時的対象を抽出する方法、およびトピック遷移構造を通時的対象を用いて分析する方法について述べる [62][38]。

7.1 通時的対象

6 章にて述べたように、文書集合全体の記述内容を時間経過に沿って確認した際に、変化を確認できる文書集合と、変化を確認しにくい文書集合が存在する。このことは、1.1.2 節に記した、

1. さまざま対象について時間の経過にかかわらず記述を行った文書集合
2. 時間の経過に沿って同一の対象について議論を行った文書集合

の違いがひとつの要因であると考えられる。

新たに、植物を観察して記録した文書が大量に存在し、その中に、以下のような記述が存在する例を考え、検討を行う。

(A) 水仙が咲いていた
(B) 梅のつぼみが大きくなった
(C) 梅の花が咲いた
(D) 梅の小さな青い実は毒だ

(A)(B)(C)(D) のうち、「水仙」と「梅」という別々の植物について記した (A)(B) は、図 7.1 のように、異なる対象への行為を記述した文書の集合（以下、多様性記述型とする）である。これに対し、「梅」について別の時刻の観察結果を記した (B)(C) は、図 7.2 のように、同一とみなせる対象への異なる時刻の行為を記した文書の集合（以下、変化記述型とする）である。なお、(B)(C) に対する (D) のように、同一とみなせる対象についての記述でも、時間の経過が意味を

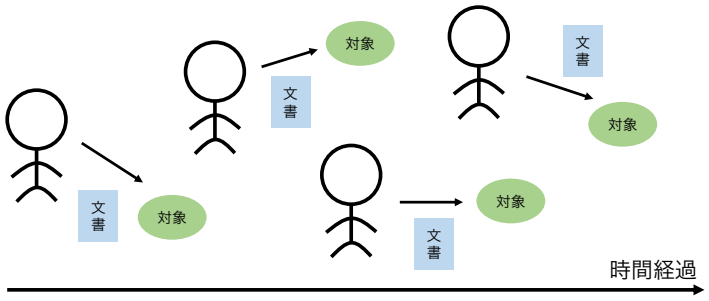


図 7.1 異なる対象への記述からなる文書集合（多様性記述型）

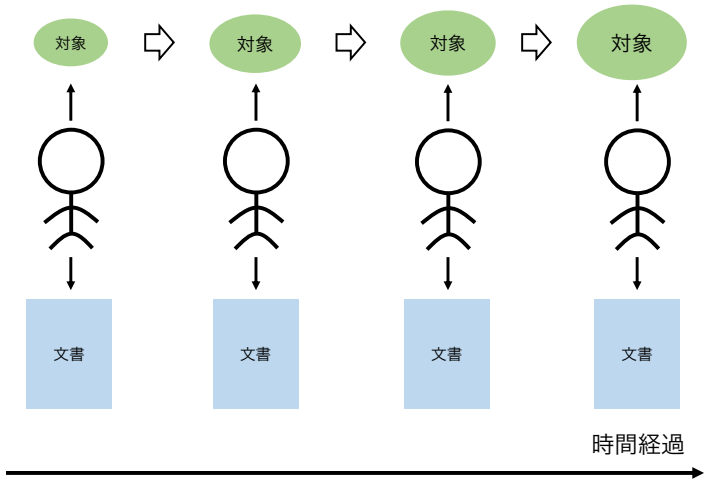


図 7.2 関連する対象への記述からなる文書集合（変化記述型）

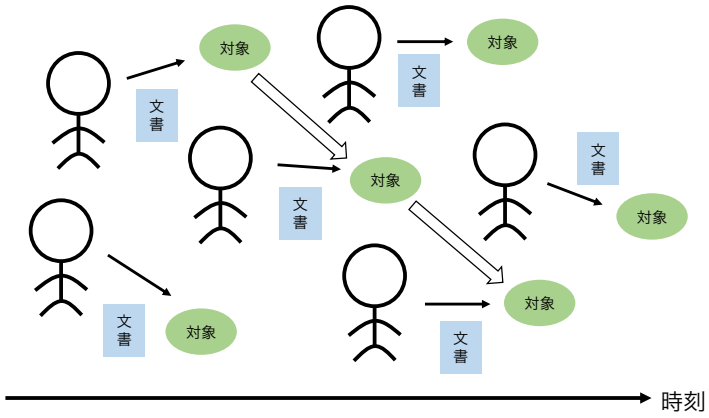


図 7.3 一般的な文書集合（多様性記述型と変化記述型が混在）

なさない場合、多様性記述型であるとみなすことにする。

以下、この「同一とみなせる対象」を、時間経過を通して扱う対象であることから、「通時的対象」と呼ぶことにする。

変化記述型の文書集合では、通時的対象への異なる時刻の行為を記述するため、文書集合から時間の経過を読み取れる。一方、多様性記述型の文書集合からは、時間の経過を読み取ることが

できない。このように、通時的対象の有無が文書集合から時間の経過を読み取れるかを決定する。すなわたち、「何かが変化したということは、なにか変化しないものがあることが必須である。」[74] に示される「なにか変化しないもの」の役割を通時的対象が果たしているといえる。

次に、植物を観察して記された文書集合をもとに、新たに植物を育てることにしたとしよう。多様性記述型の文書集合からは、どのような種類の植物があるかの情報を得ることができ、育てる植物の種類を決めることに役立つ。変化記述型文書集合からは、どのように植物が育っていくかの情報を得ることができ、時間を追って植物を育てる途中で行う行動の決定に役立つ。このように、多様性記述型の文書集合と変化記述型の文書集合からは、得られる情報の性質が異なる。

一般的に、文書をひとつの集合として捉えると、「(A)(B)(D) または (B)(C)」のように、同一の対象に関する記述か否かの明確な分類のもとに文書が蓄積されていることは少なく、図 7.3 のように多様性記述型と変化記述型が混在していると考えられる。そこで、文書集合から変化記述型となっている部分集合を見つけることを目的とし、変記述型の文書集合の核となる通時的対象の抽出と利用について、本章では検討を行う。

なお、多様性記述型の文書集合から得られた情報を、通時的対象への連続した行為へと変換し、変化記述型の結果を生み出すことが、「知的な振る舞い」であると考えられる。

7.2 通時的対象の抽出手法

7.2.1 変化記述型文書集合と単語への言及の類似率

文書の作成者が同一の対象への記述を行う際に、変化記述型（図 7.2）の文書集合では、作成者が、記述対象が以前と同じ対象であることを認識して記述を行う。文書の読者の立場から考えても、対象についての記述がある程度類似していなければ、対象が同一であると認識することは難しい。そのため、文書に含まれる同一の対象への言及内容にはある程度の類似性があると考えられる（図 7.4)。一方、多様性記述型の文書集合（図 7.1）では、以前の対象と同じかを意識せずに記述が行われるため、同じ対象を扱う文書において、対象への言及内容に類似性があるか否かは、一概に判断できない。

また、対象への言及は、その単語と共起する単語により行われると考えることができる。これを踏まえ、異なる時刻に記述された文書 A と B において、対象を表現する単語と共起する単語が類似性を持てば、文書 A と B は変化記述型の文書集合を形成すると考えられる。

そこで、対象が 1 つの単語で表現されると仮定し、単語への言及の類似率（以下、言及類似率）を求め、言及類似率が高い単語が通時的対象を示すものと仮定し、通時的対象の抽出を試みる。

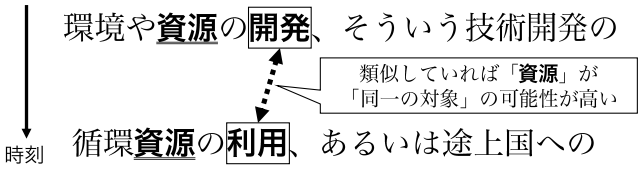


図 7.4 同一の対象への記述判断例

7.2.2　言及類似率の計算

言及類似率は、以下のように計算する。まず、着目単語 w_a を定め、ある文書 D_i において単語 w_a と共起する単語のひとつを単語 w_b とする。次に、文書 D_i とは異なる文書 D_j において単語 w_a と共起する単語 w_c に対して、単語 w_b と類似するかの判定を行う。図 7.4 の例では、「資源」が w_a に、「開発」が w_b、「利用」が w_c にあたる。さらに、単語 w_b と単語 w_c が文書集合の中の異なる時刻で使われているかの判定を行う。これを繰り返し、単語 w_a と共起するすべての単語 w_b について、その他の文書で単語 w_a と共起する単語 w_c が、異なる時刻で類似する割合を、言及類似率として求める。

処理対象とする文書からの単語の取得は、MeCab[29] による形態素解析の結果を用い、形態素解析後に名詞として得られた単語を選択することにより行った*。また、単語 w_a と 単語 w_b が共起するとは、形態素解析により得られた名詞の単語列において、w_a と w_b が 5 単語以内であることを指すものとした。

単語が類似するかの判定は、日本語版 Wikipedia に含まれる単語を Word2vec [33] によりベクトル化した Wikipedia Entity Vectors [1] から、全単語を 300 次元のベクトルとして学習済みのモデルを用い、単語 w_b、w_c を表現するベクトル $\overrightarrow{w_b}$、$\cdot\overrightarrow{w_c}$ を用いて類似度を

$$\mathrm{sim}(w_b, w_c) = \begin{cases} 0 & (w_b = w_c) \\ \dfrac{\overrightarrow{w_b} \cdot \overrightarrow{w_c}}{|\overrightarrow{w_b}||\overrightarrow{w_c}|} & (w_b \neq w_c) \end{cases} \tag{7.1}$$

として求め、$\mathrm{sim}(\overrightarrow{w_b}, \overrightarrow{w_c}) \geq 0.5$ となる単語 w_b、w_c を類似する言及とした。なお、処理対象の文書集合に存在するが、Wikipedia Entity Vectors に存在しない単語は、言及類似率の計算対象外とした。

単語 w_b と単語 w_c が異なる時刻で使われているかの判定は、単語 w_a と共起して単語 w_b が出現する時刻、同様に単語 w_a と共起して単語 w_c が出現する時刻の分布に着目して行ない、相互の第 2 四分位範囲から第 3 四分位範囲に重なりがなければ、異なる時刻で使われているものとした。

* 名詞のうち、「名詞-形容動詞語幹」（「必要」など「〜ない」となる単語）や「名詞-サ変接続」（「実験」「試験」など「〜する」となる単語）は、通時的対象となる可能性が低いと考えられるため除外した。

小型人工衛星設計議事録	
期間	2000 年 1 月 5 日〜2002 年 12 月 12 日
文書数	398
異なり単語数	7,877

環境省中央環境審議会地球環境部会議事録	
期間	2001 年 2 月 16 日〜2012 年 10 月 24 日
文書数	5,910（発言）
異なり単語数	12,991

ツイート集合（2014 年前半の「人工知能」検索結果）	
期間	2013 年 12 月 25 日〜2014 年 6 月 6 日
文書数	43,862（収集データの $\frac{1}{3}$）
異なり単語数	22,251

表 7.1 対象文書集合の概要

7.3 通時的対象の抽出実験

7.3.1 抽出対象とする時系列文書集合

5.1.1 節に述べた小型人工衛星設計議事録、6.1.2 節の中央環境審議会地球環境部会議事録、および 6.1.3 節のツイート集合（2014 年前半の「人工知能」検索結果）を対象とし、通時的対象の抽出を行った。各文書集合の作成期間、文書数、異なり単語数を表 7.1 に示す。

7.3.2 抽出結果上位語の評価

表 7.4、7.3、7.5 に、各文書集合から得た言及類似率の上位 10 語と、出現頻度が上位 10 語を比較のために示す。

抽出結果を確認するために、抽出結果の上位の各単語について、単語とその周辺の記述を元の文書の内容を確認し、各単語が通時的対象であるか、各単語が出現する箇所の記述を確認し、Discounted Cumulative Gain（以下、DCG）[27] を求めるための評価値を付与した。

DCG_k は、文書検索システムの評価などに用いられる値であり、評価対象となる結果として得られたうちの上位 k 個を用い、以下の手順により求めることが出来る。ここでは $k = 10$ とした。

$$\mathrm{DCG}_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2 i} \tag{7.2}$$

rel_i は i 番目の結果の関連性評価値である。本実験では、言及類似率の大きい単語が通時的対

象であるかを評価することを目的とし、関連性評価値を表 7.2 のように定めた。

評価値	評価概要
3	評価語を対象とした変化記述がなされている
2	評価語が複数の対象を表し、複数の変化記述がある
1	評価語が複数の対象を表し、変化記述と多様性記述が混在する
0	評価語が記述の対象ではない、または変化記述がない

表 7.2 通時的対象かを判断するために用いた関連性評価値

なお、出現数が多い単語では、文書集合の中で出現箇所が数千以上にのぼる。そのため、評価対象の各単語について、出現箇所をランダムに 20 か所選択し、記述を確認することにより評価値を定めた。この際、変化記述型文書集合の対象は、図 7.3 のように時間経過に沿って出現すると考えられるため、文書集合を作成時刻順に 5 つのグループに分け、各グループから 4 文書ずつを選択した。

	言及類似率			出現頻度		
	単語	割合	$\boldsymbol{rel_i}$	単語	割合	$\boldsymbol{rel_i}$
1	熱	0.5733	3	XI	0.0095	3
2	構造	0.5726	1	電源	0.0090	0
3	他	0.5653	0	アンテナ	0.0082	3
4	CUBE	0.5642	3	月	0.0077	0
5	方法	0.5633	0	基板	0.0072	3
6	周期	0.5628	2	データ	0.0067	2
7	クリーンルーム	0.5627	3	電圧	0.0062	1
8	ボード	0.5625	3	太陽電池	0.0061	3
9	画像	0.5619	3	温度	0.0060	2
10	真空	0.5613	2	地上	0.0058	2

表 7.3 上位 10 語と評価（小型人工衛星設計議事録）

また、関連性評価値が 1 以上であれば、変化記述型の文書集合を探す手がかりとすることが出来ることから、表 7.4、7.3、7.5 において $rel_i \geq 1$ である単語の割合を、精度として求めた。

表 7.6 に、各文書集合における、言及類似率と出現頻度がそれぞれ上位 10 語の単語について、精度と DCG_{10} の値を示す。

7.3.3　言及の類似度の分布

言及類似率の計算を行う着目単語 w_a に対する言及単語 w_b について、異なる時刻で行われる言及 w_c のうち類似度が最大となるものについて、その類似度ごとの w_b の数の分布を、図 7.5 に示す。

小型人工衛星設計議事録において言及類似率が最も高かった単語「熱」と、出現頻度が最も高

	言及類似率			出現頻度		
	単語	割合	rel_i	単語	割合	rel_i
1	法律	0.5969	3	委員	0.0101	0
2	海外	0.5916	3	環境	0.0095	2
3	ルール	0.5910	3	エネルギ	0.0088	2
4	大気	0.5908	2	日本	0.0073	3
5	資源	0.5899	2	資料	0.0072	0
6	目的	0.5897	1	目標	0.0065	2
7	効率	0.5895	2	地球	0.0047	2
8	条件	0.5886	0	技術	0.0047	2
9	家庭	0.5883	3	部会	0.0044	0
10	公共	0.5883	3	制度	0.0044	2

表 7.4 上位 10 語と評価（地球環境部会議事録）

	言及類似率			出現頻度		
	単語	割合	rel_i	単語	割合	rel_i
1	人間	0.5850	2	人工知能	0.0684	1
2	ロボット	0.5830	0	表紙	0.0375	2
3	自分	0.5801	1	人工知能学会	0.0320	0
4	人	0.5776	1	女性	0.0309	0
5	根本	0.5741	0	ロボット	0.0292	0
6	社会	0.5726	0	男	0.0225	0
7	形	0.5719	1	家事	0.0220	0
8	機械	0.5708	1	気持ち	0.0218	0
9	学会	0.5705	3	まとめ	0.0214	0
10	人類	0.5701	1	NAVER	0.0201	0

表 7.5 上位 10 語と評価（ツイート集合）

	地球環境部会議事録		小型人工衛星設計議事録		ツイート集合	
手法	P	DCG_{10}	P	DCG_{10}	P	DCG_{10}
言及類似率	0.9	12.70	0.8	9.891	0.7	5.068
出現頻度	0.7	7.517	0.8	9.548	0.2	3.000

表 7.6 文書集合ごとの精度 (P) と DCG_{10}

かった単語「XI（設計対象の衛星の名称)」について、言及している単語の言及類似率の分布を確認すると、「熱」では類似率 0.6〜0.7 の単語が最も多く、「XI」では類似率 0.5〜0.6 の単語が最も多い。

地球環境部会議事録における「法律」と「委員」を比較すると、「法律」の方がわずかだが類似度の高い言及の割合が多いことが分かる。議事録中でも、「法律」は「法律の施行」「法律の適用可能生」「法律体系全般」などの言及が行われているのに対し、「委員」は「(役職や個人名を示

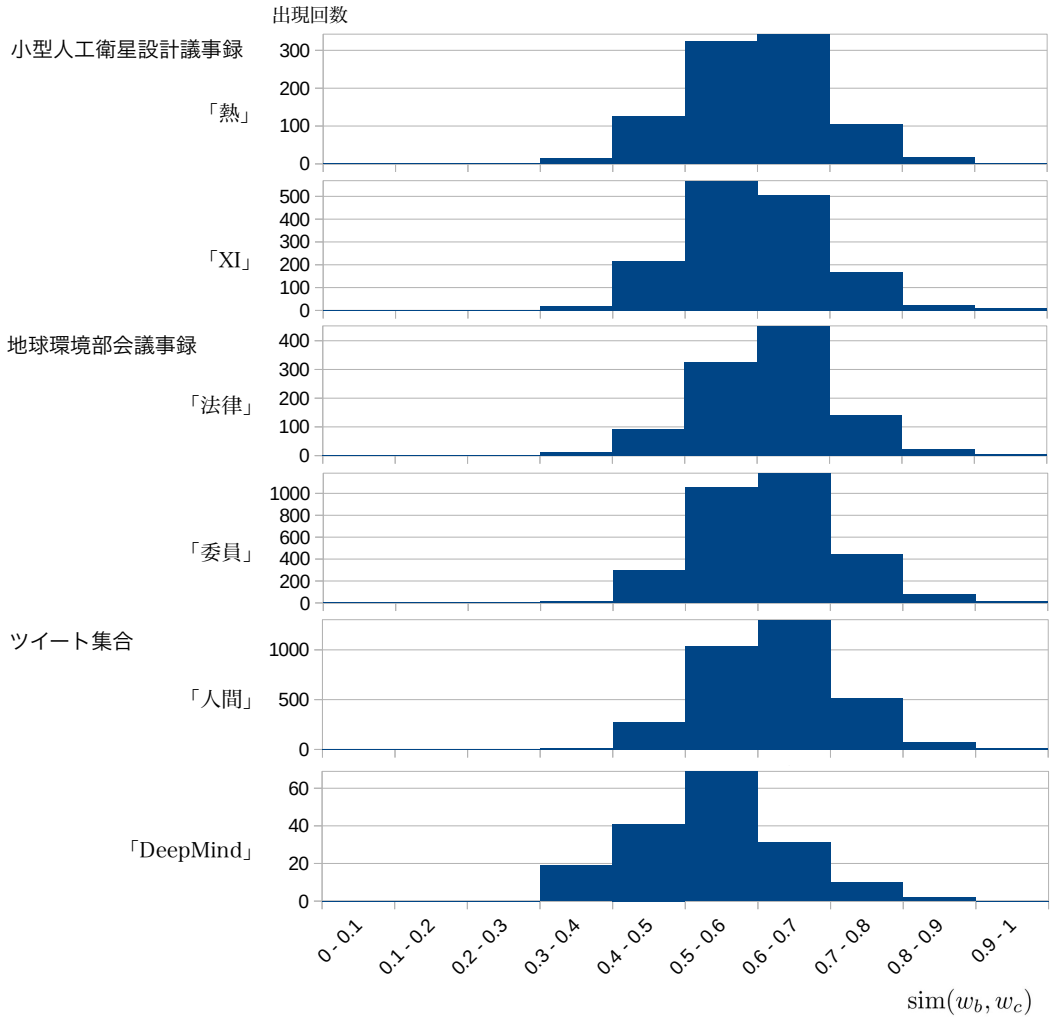


図 7.5 着目語に対する異なる言及 w_b、w_c の $\mathrm{sim}(w_b, w_c)$ の最大値の分布 (a)「法律」、(b)「委員」(地球環境部会議事録)、(c)「人間」、(d)「DeepMind」(「人工知能」を含むツイート)

して）○○委員」と出現することが多く、言及内容は会議に関連した範疇ではあるが、多岐にわたる。

ツイート集合における「人間」は、地球環境部会議事録の「法律」「委員」と類似した分布を示す。一方、「DeepMind」には、類似度の低い言及が多いことが分かる。「DeepMind」を含むツイートを確認すると、「DeepMind が Google に買収された」ことについて、外部の記事の紹介ではなく、各ツイートそれぞれの感想を述べており、同じ事象について関連性なく記述が行われた結果であると考えられる。

7.3.4　考察

抽出実験の結果、上位 10 語のうち、変化記述型の文書の通時的対象と考えられる単語は、言及類似率を用いたほうが DCG_{10}、精度とも値が高く、目標とした通時的対象の抽出が行えている。

小型人工衛星の設計議事録においては、精度は言及類似率でも出現頻度でも同じであり、DCG_{10} の値の差も小さかった。この理由として、小型人工衛星設計議事録では、人工衛星の部品名称を示す単語の出現頻度が高く（「電源」「アンテナ」など)、設計対象の部品であることから変化記述型の記述対象となっていたことが挙げられる。

ツイート集合では、言及類似率と出現頻度で得られた上位 10 語の精度（P）の差が大きい。これは、同一のまとめ記事・ニュース記事のタイトルを紹介するツイートが非常に多数あり、出現頻度の高い単語はこれらのタイトルに由来するためである。

7.4　通時的対象の抽出における課題と展望

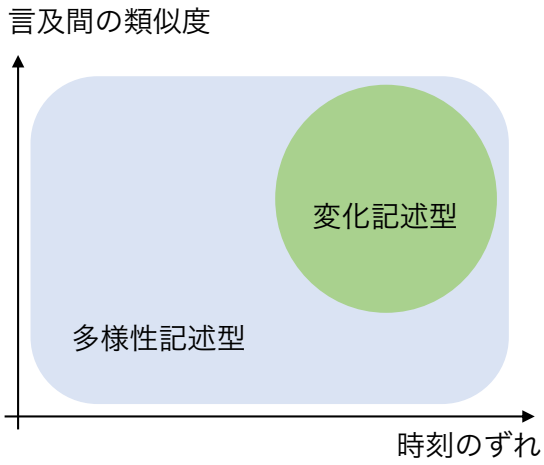


図 7.6 多様性記述型・変化記述型文書集合の分布例

通時的対象の抽出における課題と、今後の展望を述べる。

変化記述型の文書集合には、「異なる時刻において共起語が類似する単語が存在する」ことを仮定して実験を行った。しかし、多様性記述型の文書集合でも、別の時刻の同一の対象について以前とは無関係に記述を行うことがあることから、得られた結果は、両方の記述形式が混ざったものとなる。両形式の文書集合を、対象への言及内容の類似度と記述時刻のずれを軸として整理すると、図 7.6 のように重なる。「同一とみなせる対象」を文書集合から見つけるためには、この重なりを区別する手法が必要である。

本稿では、「対象がある単語で表される」としたが、同一の単語が異なる意味で使われることは多々ある。これに対応するために、対象単語を含む記述の全てから文書をランダムに選択して評価を行うのではなく、評価の対象とする単語への言及類似度が高い部分ごとに分割することにより、変化記述型の文書集合を選択できるか、検討を行う。

また、実験結果を確認するために、言及類似率上位の単語と出現頻率が上位の単語との比較を行ったが、単語 w_a とそれぞれ異なる文書で共起する単語 w_b、w_c は、単語の共起関係をネット

ワークと考えた際、単語 w_a を媒介として接続されことから、媒介中心性と言及類似率に類似が見られないか、比較を行いたい。

次に、提案手法における単語間の類似度の計算のために Wikipedia Entity Vectors として配布されている Wikipedia の記述内容をもとにした単語のベクトルデータを用いた。このため、実験対象の文書集合には出現するが Wikipedia Entity Vectors に含まれない単語の類似度計算は行えない。計算対象とする文書集合も含めて単語のベクトル表現を求めることにより、文書中のすべての単語について、類似度計算を行うことができる。また、類似度計算が正確に行えているかを比較、評価することも必要である。

7.5　本章のまとめ

本章では、文書集合には多様性記述型と変化記述型の 2 通りが考えられることを述べ、変化記述型の記述の「対象」を抽出するため、言及類似率を提案し、言及類似率が高い単語を「通時的対象」とした。さらに、提案手法を 3 つの形式が異なる文書集合に適用し、通時的対象の抽出が行えていることを確認した。

さらに、実際の文書集合は多様性記述型、変化記述型と 2 つに分かれているのではなく、その多くが多様性記述型の文書集合であり、そのなかに通時的対象の時間の経過に沿った変化を述べた変化記述型の文書集合が部分集合として含まれていると考えられることを示した。

第 8 章

時間経過に沿った変化の工学的利用

本論文ではここまでに、時間の経過に沿って蓄積された文書集合を時間の経過に着目して扱うシステムを提案し、いくつかの文書集合への適用事例を述べた後、文書集合の性質について検討と時間の経過に着目した変化を取り出すための通時的対象の提案を行った。

本章では、工学的な利用に的を絞り、時間の経過に着目した変化がどのように利用可能か、またそのためにはどのように文書集合を蓄積するのが適切か、整理と検討を行う。

8.1　時間経過に沿った変化から得られる知識

本論文では、文書集合の理解に時間経過の視点を導入し、時間の経過に沿った変化を述べた変化記述型の文書集合に着目してきた。

1.1.2 節、7.1 節に述べたように、変化記述型の文書集合からは、設計などの行為を時間の経過に沿って進める際に、「どのように作業を進めるか」の決定を行う際に役立つ。すなわち、文書集合を時間経過に着目して扱うことにより、文書集合から、時間に沿って進める行動の「手順」と行動の「根拠」の 2 種類の知識を得られる可能性がある。

8.1.1　時間経過に沿った手順

5.4.3 節に述べたように、4 章で構築したトピック遷移構造の抽出・再構成システムにより、システムの利用者が指定した対象への連続あるいは並行して行われる作業を抽出することができる。例を図 8.1、図 8.2 に示す。

この機能は、トピック遷移構造から、システムの利用者が指定した単語を指定確率以上で含むトピックを選択し、ラベルとして指定語と共起するサ変名詞を表示することにより実現した。表示には、トピック間のリンクや同時刻に存在するトピックのレイアウトが見やすく行われることから、4.4.1 節に述べた静的なグラフ提示を用いた。

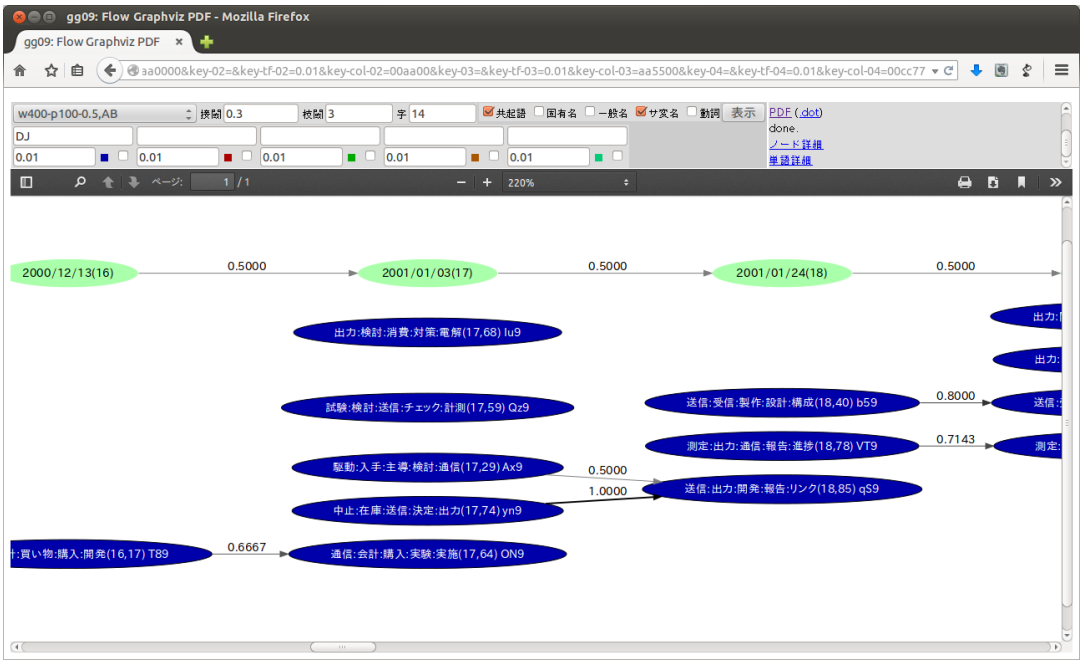


図 8.1 「DJ」と共起するサ変名詞（一部）・表 5.4 の設定による（図 5.13 再掲）

図 8.2 「西無線」と共起するサ変名詞（一部）・表 5.4 の設定による

「DJ」について、図 8.1 で得られたトピック同士の関係と、各トピックに含まれる文書を確認することにより、「購入を検討していたこと」「製造中止になり在庫の確認が必要であること」「出力に関して検討していたこと」が同じ時期に並行して議論され、さらに「出力の測定」は継続して行われていることを確認できた。

同様に「西無線」について、図 8.2 と各トピックに含まれる文書をあわせて確認し、「製作仕様の交渉」「シールドを提示」が並行して行われ、続いて、「他の無線系統との関係」「電力と出力に関する検討」「アンテナの切り替えに関する検討」、さらに「CW ジェネレータのプログラムの改良」行われていることを確認できた。

このように、時間経過に着目して文書集合を処理することにより、複数の作業が並行して行わ

れることも含めて、対象へ行われている作業を時間経過に沿った手順として抽出することが可能である。抽出した手順を用いることにより、作業過程の検証や、あらたな対象への作業手順の検討を行うことができる。

8.1.2 時間経過の中に含まれる根拠

5.4.2 節に述べたように、トピック遷移構造の抽出・再構成システムにより、時間経過に沿って発生した変化の根拠、たとえばある部品からある部品への変更の理由を、文書集合から見つけることが可能である。

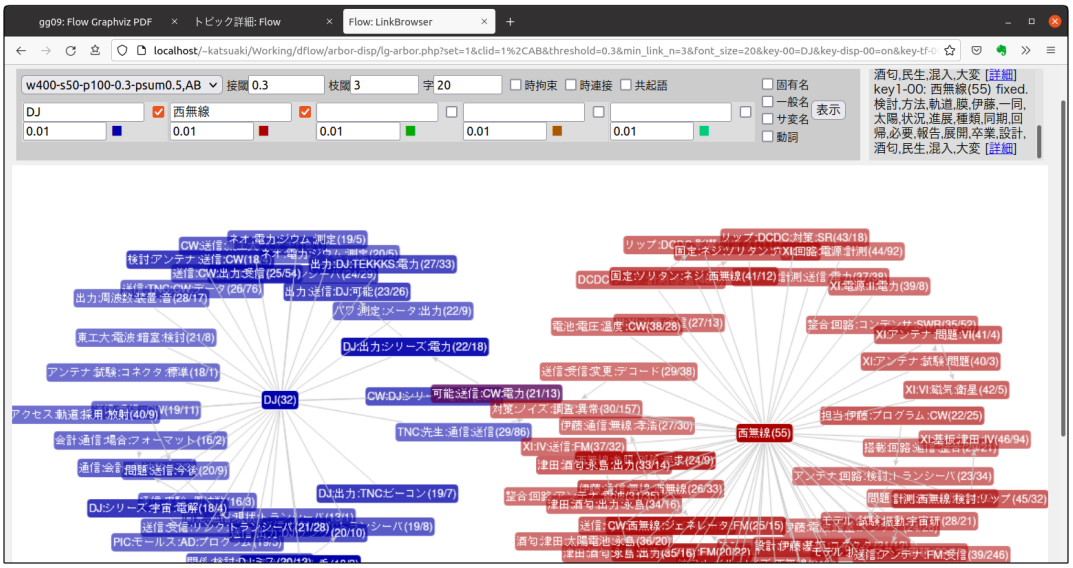


図 8.3 「DJ」と「西無線」によるトピック遷移構造の再構成結果

図 8.3 は、小型人工衛星設計議事録から抽出したトピック遷移構造に対し、「DJ」と「西無線」の 2 種類の無線機をキーワードに設定して再構成を行った結果を、力学的モデルによる提示により表示した例である。力学的モデルによる提示は、トピック遷移構造に含まれる各トピックや、システム利用者がしていた着目対象などの距離を 2 次元空間上で把握することに適している。

図 8.3 において両者に共通するトピックに含まれる文書から、西無線にどのような仕様で無線機の設計を依頼するかなど、西無線への無線機製作を依頼する際の検討事項の記述を発見できた。

このように、トピック遷移構造から、関係性を確認したい複数の対象に関連したトピックの遷移を確認することにより、変更の根拠を確認することが可能である。この根拠を確認することにより、過去の判断を再確認すること、新たに判断を変更する場合の再検討などを行うことができる。

8.2　文書集合処理の限界

ここまで、時間経過に着目することにより文書集合から手順や根拠を得られることを、あらためて述べた。一方、文書集合中に記述されていないことは、時間経過に着目しても有用な知識を得られない場合がある。

8.2.1　得られた手順の汎用性

8.1.1 節では、「抽出した手順により新たな対象への作業手順の検討を行うことができる」と述べた。しかし、例として述べた「DJ」「西無線」に関する手順は、それぞれの対象の固有の手順である。すなわち、文書集合中に記述された手順は個別の対象に関する手順であり、そのままでは、新たな対象へ適用することはできない。何らかの手法により手順を一般化した後、あらたな対象に適用できるように、再度詳細化する必要がある。

8.2.2　記述のない根拠

8.1.2 節では、トピック遷移構造を確認することにより文書集合から根拠を抽出可能であると述べた。しかし、文書集合から根拠を発見することはできず、推測するしかない場合もある。

例えば、5.4.1 節に述べた小型人工衛星設計議事録において「モデム」「TNC」の関係を確認する場合、トピック遷移構造により「モデム」「TNC」の出現状況を確認した結果、および「TNC」が「モデム」を含む機器であることを調査した結果から、「モデム」に関する検討が「TNC」に関する検討に統合されたと推測することができる。しかし、文書集合中には、「TNC はモデムを含む機器である」「モデムの検討は、以後 TNC の検討として行う」といった記述は出現せず、文書集合のみから、この推測の根拠を得ることはできない。

この推測は、本研究が提案するシステムの利用者が、「TNC はモデムを含む機器である」という知識を文書集合外部から持ち込み、「『モデム』は文書中に現れないが『TNC』はたびたび現れる」という文書集合から得られた情報と組み合わせることにより、はじめて可能となった。

もう 1 つ、例として、小型人工衛星設計議事録において、「衛星と太陽電池を接続したが太陽電池が検出されない」という事例をとりあげる。この問題の原因と解決策として、「太陽電池と衛星本体の GND を比較すると、太陽電池の GND の方がレベルが低いことが原因であり、太陽電池の GND を全体の GND とする」という記述が文書集合中に存在する。

この解決策の記述より前において、文書集合中に問題の根拠となる記述が含まれないか確認を試みると、トピック遷移構造より、「太陽電池」「衛星本体（基盤など)」それぞれと「GND」に関する記述を見つけることはできる。しかし、「太陽電池の GND」と「衛星本体（基盤など）の

GND」の電位の比較が必要か否かは、記述されていない。

「記述されていないことが問題の兆候である」と捉えることは可能だが、これは、「複数の部分構造間の GND の比較が必要である」という知識を持っていて初めて可能なことである。

このように、文書集合中に記述されていない根拠は、文書集合中から見つけることはできない。

8.2.3 先行研究における解決策

前節では、文書集合を時系列に着目して得た情報の限界として、手順が一般化できないこと、記述されていない根拠は見つけられないことを述べた。そこで、先行する研究ではこれらの課題の解決をどのように試みているかを確認する。

まず、手順の抽出に関する研究として、2.7.1 節に述べたプロセスマイニングが挙げられる。プロセスマイニングでは、「現状の（情報）システムですでに利用可能なイベントログ」[42] を手順抽出の対象とする。例えば、ソフトウェアの開発プロセスにおける研究では、バージョン管理システムのログを用いた研究 [35]、プロジェクト管理ツールのログを用いた研究 [31] などが行われている。イベントログからのマイニングの過程において、複数回出現するイベントの系列を手順として扱うことにより、手順の一般化を行うこともできる。

次に、根拠を扱う研究として、2.5.2 節に述べた設計支援システムを挙げる。例えば gIBIS[11] は、議論と並行して、議論の内容を 何らかの判断が必要な問題である Issue、解決案である Position、解決案に関する情報である Argument に分類し、それらの関係を画面上で整理、記録しながら設計と議論をすすめるシステムである。また、設計支援システムでは、Issue に基づく方式以外に、2.5.2 節に述べた設計対象のモノと人間の設計活動を統合的に扱う設計支援を行う研究 [61]、2.4 節に述べたオントロジーを中心として人工物の設計・運用中を行うシステム [57] なども研究が行われている。

8.2.4 先行研究の課題

先行する研究では、記録を正確に行うために情報システムを利用することを前提としている。プロセスマイニングは情報システムのイベントログの存在を前提とし、設計支援システムは、設計過程の記録のために記録用のシステムが使用されることを前提とする。とくに、設計支援システムは、研究の一環として提案されたシステムが記録に用いられる必要がある。しかし、研究用途のシステムよりも一般的であると考えられる、UML や IDEF0 などのモデリング方式であっても、認知されている度合いは 2 割程度 [69] である。そのため、これらの研究により提案されているシステムが研究の場以外で活用され、実際に記録を行うために使用される可能性は低いと考えられる。

8.2.5　本研究の利点

先行研究が手順や根拠を扱うために情報システムを用いた記録を必要とするのに対し、本研究では、議事録などの形で蓄積された文書集合を対象として時間経過に沿った変化に着目することで手順や根拠を扱っており、情報を蓄積する際に特別な仕組みを必要としない。

また、本研究では、文書集合から時間経過に沿ったトピックの遷移を抽出し表示するだけではなく、表示を操作するシステムを設け、システム利用者の操作に応じて再構成を行う。そのため、システムの利用者が文書集合から得られる情報では自身の目的達成に不十分であると感じた際に、文書集合の外部の情報を収集し、文書集合に記されていない事柄を補完することが可能である。

8.3　文書集合の形成形式の検討

ここまでに、すでに存在する文書集合を対象に時間の経過に着目したトピック遷移構造の抽出・再構成を行うことにより、手順や根拠を抽出することが可能であることを示した。しかし、どのような文書集合にでも同様な可能なわけではなく、文書集合に記述されていないことは扱えず、6.5.4 節に述べたように、記述の対象が、複数の文書にまたがり共通して存在しその変化が記述されている通時的対象が必要である。本節では、どのように文書を作成し蓄積すれば、本研究の手法にとって有効な記述がなされるかについて、検討を行う。

8.3.1　文書集合ごとの蓄積形態の確認

これまでにとりあげた文書集合の蓄積形態の違い

小型人工衛星設計議事録では、7.3.4 節に述べたように頻出単語も通時的対象となっていた。そのため、文書の作成・蓄積過程において、小型人工衛星設計議事録と同様の形態をとれば、通時的対象を多く含む文書集合を生成することが可能である。

小型人工衛星設計議事録に含まれる各文書は、その時点までの設計の結果を確認する会議の議事録として作成されている。すなわち、それぞれの文書は図 8.4（図 3.12 を再掲）のように、対象へ人間 A が働きかけを行う様子を観察した人間 B により作成され、それらを人間 C が解釈、利用する。

会議の議事録として文書に記されるのは、それ以前の行為をまとめた結果である。3.1.1 節に述べた「作曲」と同様、外部に提示されることを意図しない過程を積み重ねた結果を、外部に提示するためにまとめたものが、文書として記録されている。

一方、審議会議事録とツイート集合においては、「発言すること」「ツイートすること」といっ

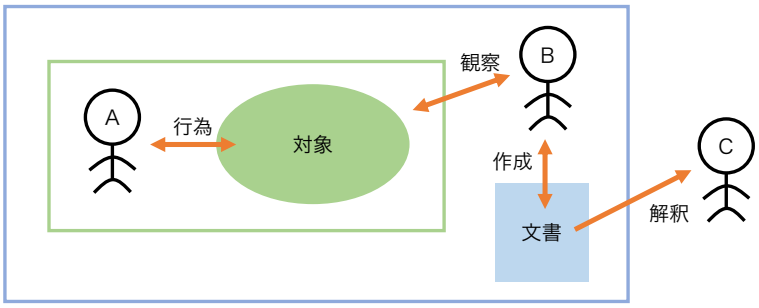


図 8.4 設計議事録における人間・文書・対象の関係（図 3.12 再掲）

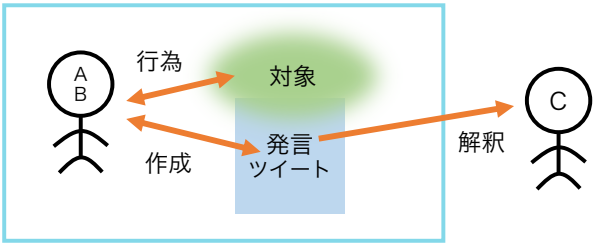


図 8.5 審議会議事録・ツイート集合における人間と文書・対象の関係

た文書の作成自体が対象への行為であり、対象への行為と文書の作成（発言）を行う主体が同一である（図 8.5）。また、発言中に自らの発言に言及するなど、発言自体が対象である状況も考えられる。このように、文書の作成時点において、行為と文書の作成、対象と文書、どちらも明確には分かれていない。また、このように行為と文書の作成が同時に行われる状態は、3.1.1 節に述べた「演奏」と同様の形態であり、個別の行為そのものが、外部に提示されることを意図したものである。

そのため、これらを記録した文書を解釈する人間 C にとって、具体的な対象を読み取ることができない文書が作成され、文書集合として蓄積されると考えられる。

蓄積形態間の関係・位置づけの関係

2 種類の文書集合の蓄積形態、図 8.4 と図 8.5 は、図 8.6 のような関係にあると考えることができる。

設計会議など（以下、形式 A）の形式で作成された文書が、審議会など（以下、形式 B）での解釈対象となる。形式 B では、解釈対象の文書をもとに、議論の対象が何であるかの検討、議論対象についての検討が行われる。すなわち、形式 B は、「すでにある程度の検討が行われた結果」を再確認する場合であり、「新しい結果」を生み出す場ではない。形式 B で表出された情報をもとに、形式 A により作成された結果が承認されるか、形式 A のような外部で新たな検討が行われるかのいずれかである。形式 A では、形式 B のような外部の場の検討結果を取り込み、現在の方針を継続するか、あらたな方針を検討するかなどの判断を行い、継続して作業を進めていく。

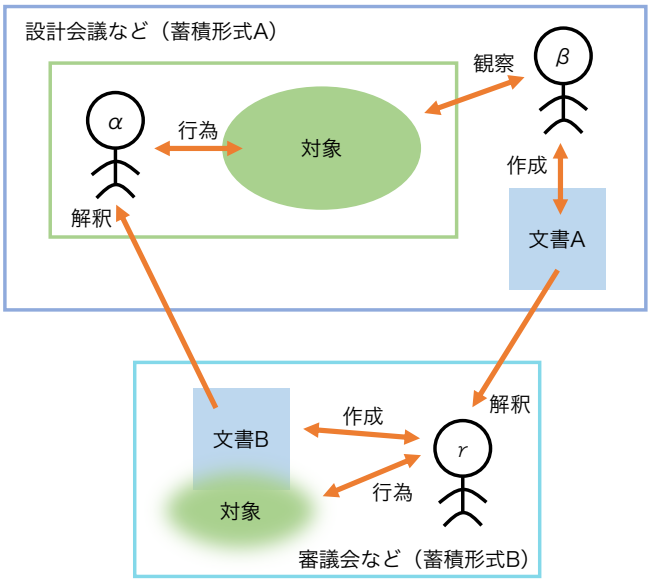


図 8.6 設計議事録と審議会議事録の関係

時間の経過に着目した処理に適した文書集合の蓄積形態

以上を踏まえると、図 8.4 のように、対象の操作など何らかの行為を行う場と、それらを観察し文書を作成する場を分けることにより、通時的対象が存在し、時間の経過に伴う変化の抽出により手順や根拠の抽出が行いやすい文書集合を形成できると考えられる。行為と文書の作成を分ける際には、行為を行う主体と文書作成を行う主体を別の人間とする、あるいは対象への行為と文書作成を空間的・時間的に別の場所に分ける、などの手法がとれるだろう。この分離により、通時的対象と通時的対象に関する記述が文書集合により多く記録されることが期待される。

8.4　本章のまとめ

本章では、本研究で提案した文書集合を時間経過に沿って処理する手法によって得られる、手順や根拠といった工学的に利用可能な知識について述べた。先行する研究では、設計などの実施途中にその過程を記録するシステムを導入することにより手順や根拠の取得を目指したが、本研究では、すでに記録された文書集合から手順や根拠の取得を行うため、記録を行うための情報システムは必要ない。

しかし、時間の経過に沿った変化を抽出するためには、複数の文書が共通した通時的対象を含むなど、抽出に適した文書集合であることが望ましい。そこで本章では、時間経過に沿った変化の抽出に適した文書集合を蓄積するために適した、記述の対象、記述を行う人間、記述される文書の関係を検討し、対象への行為と文書の作成とを分けて行う形式が適切であることも述べた。

第 9 章

結論

本研究は、文書集合の理解に時間経過の視点を導入することを目的とし、時間の経過に沿って蓄積された文書集合の理解を支援するシステムの構築、文書集合の内容の時間経過との関係の考察を行った。本章では、本論文のまとめと学術的貢献、今後の展望について述べる。

9.1　本論文のまとめ

第 2 章では、本研究と関連する研究について、時間の経過に着目して情報を扱う先行研究について述べ、本研究の位置づけを行った。まず、時系列に沿った情報の可視化についてとりあげた後、文書集合から時間経過に沿ってトピック抽出を行い俯瞰表示を行う研究、時系列データから因果関係の抽出しベイジアンネットワークを構築する研究、時間経過に沿って知識を伝達する過程の研究などを取り上げた。続いてこれらの研究を、着目する時間経過の長さ、処理対象とするデータに対する事前知識の必要性により整理した。最後に、本研究では、長い時間経過を扱い、処理対象とするデータ（文書集合）への事前知識を必要としないことを述べた。

第 3 章では、文書集合の蓄積形態に関してその形式を議論し、以後の章で用いる基本的概念の整理を行った。まず、時間経過の長さについて「作曲の繰り返し」と「演奏」を例に挙げ、本研究で扱う「長い時間経過」が、作曲間の内容の変化に相当することを述べた。続いて、文書と記述対象の関係について検討し、先行研究では文書集合からの出来事に関する文書記述間の関係抽出を主目的とするが、文書集合には、出来事以外に時間経過に伴い変化する対象を記述したものがあること、また、対象の変化の経過を記述する場合、結果のみを記述する場合が考えられることを述べた。さらに、文書を記述する人間、対象の変化を引き起こす人間と文書の関係についても整理を行った。

第 4 章では、文書集合に含まれる記述内容を時間経過に沿ったトピック遷移構造として抽出するシステム、文書集合の利用者の操作に応じてトピック遷移構造の再構成を行うシステムを構築し、システムの各機能について述べた。トピック遷移構造の抽出は、文書集合を文書作成時刻に

基づいて複数の文書グループへ分割し、長い文書を一定の長さの断片へ分割した後、文書グループごとに PLSI を適用することにより行う。この際、古いトピックに属する文書は忘却させつつ、以前の文書グループから、次の文書グループへ文書を追加することにより、文書グループ間の連続性をもたせた。トピック遷移構造の再構成は、本システムの利用者が指定する単語と出現確率に基づき、トピック遷移構造から条件に合致する部分構造を選択することにより行った。また、トピック遷移構造の俯瞰表示やアニメーション表示を行う他、前述した単語による再構成、再構成時に指定する単語の選択支援を行う仕組みなど、トピック遷移構造をインタラクティブに操作する仕組みを設けた。

第 5 章では、第 4 章で述べたシステムを小型人工衛星の設計議事録に適用した結果を述べた。まず、システムによりトピック遷移構造および俯瞰表示が行えることを確認した。その後、部品名称を指定してトピック遷移構造の再構成を行うことにより、関連する部品の名称を文書集合中から見つけられることを述べた。続いて、部品名称の出現確率の推移から、別の部品の一部として設計されていく様子を把握可能であることを述べた。また、途中で置き換えられた複数の部品の名称を指定してトピック遷移構造を再構成することにより、置き換えの経緯を記した議事録中の記述を見つけ出すなど、設計の根拠を把握できることも示した。

第 6 章では、第 4 章のシステムを政府審議会の議事録、Twitter から収集したツイートに適用し、時間経過に沿った記述内容の把握を試みた。その結果、俯瞰表示により異なる記述内容を分離して把握すること、再構成により内容に重複がない部分を探して俯瞰表示では見つけられなかった記述内容を把握することが可能であることを述べた。その一方、文書集合中で時間の経過に伴い新たな内容が記述されている部分を見つけることは困難であった。そこで、第 3 章の議論を踏まえて、第 5 章、第 6 章にて扱った文書集合の蓄積形態について検討を行い、記述対象が文書集合中では影響を受けない仮想的なものである、あるいは文書ごとに記述対象が異なるものであるなどの場合があり、対象が時間経過に沿って変化しないために、文書集合からは時間の経過を把握しづらいと考えられることを示した。

第 7 章では、第 6 章での議論を踏まえ、文書集合から文書間をまたがり時間経過に沿って変化する通時的対象を抽出する手法を検討し、その評価を行った。通時的対象は、文書集合中で類似するが完全には一致しない言及が異なるタイミングにおいてなされている単語により表されると仮定した。類似する言及が異なるタイミングで行われている割合を言及類似率として求め、言及類似率が高い単語と文書中の出現率が高い単語の比較を行った。この結果、言及類似率が高い単語は、通時的対象である割合が高いことを示した。

第 8 章では、トピック遷移構造の抽出・再構成システムの利用を通して得られた、文書集合から抽出した時間経過に沿った変化の工学的に利用について述べた。時間の経過に着目することにより、手順や根拠など工学的に再利用可能な情報を文書集合から取得可能なこと、その一方、文書集合に記述されていない情報は取得できないことを確認した。さらに、先行する設計根拠の獲得に関する研究と比較し、文書集合を利用する本研究の手法は、文書の作成時点で特別な情報シ

ステムなどを用いる必要がないという利点があることを述べた。また、文書集合から変化を抽出するためには、文書を蓄積する過程において、記述対象に対する行為とそれらを文書化する作業の分離が有効であることを述べた。

9.2 本論文の学術的成果

本論文では、時間経過に沿って蓄積された文書集合に対し、以下を示した。

1. 文書集合を理解するために時間経過を活用する手法があること
2. 時間経過を活用するためには、実時間に基づいた俯瞰的な提示では不十分であり、文書を理解しようとする者にあわせてインタラクティブに処理結果を操作する仕組みが必要であること
3. 文書集合において時間の経過に沿って内容が進展するためには、一定期間以上に共通する具体的な記述対象が存在すると良いこと

とくに、多くの内容を含む文書集合に対し、俯瞰的に時間経過に沿ってトピック抽出を行うだけではなく、システムの利用者の着眼点にあわせて抽出結果を再構成し、利用者の興味に関連したトピックの遷移を取り出して提示することと、着眼点の候補となる通時的対象を求める手法を提案したことは、本研究独自の成果である。

また、時間経過への着目が文書集合の理解に対して有効であるためには、複数の文書が共通した記述対象を持ち、対象への試行錯誤の過程を記録するという形態で蓄積され文書集合が形成される必要があると示せたことも、重要な成果である。

9.3 課題と今後の展望

9.3.1 トピック遷移構造抽出手法の検討

本研究では、2.8.2 節に述べたように、利用者が処理対象とする文書集合に関する知識を持つことを前提とせず、システムをインタラクティブに操作することにより文書集合を理解することを目指した。そのため、捜査対象となるトピック遷移構造の抽出には、操作システムが利用しやすい結果が得られるよう、4.3 節にて述べた本研究独自の手法を用いた。

トピック抽出および時系列文書からのトピック抽出に関する研究は、近年も新しい手法が提案されている [13][26][39] 。これらの研究では、従来通りニュース記事や学術論文などを正確に俯瞰的することを目的としているため、抽出結果の理解には事前知識が必要であるが、これらの研究の手法を、トピック遷移構造の抽出に応用することは可能であると考えられる。

また、近年の文章の処理を行う研究では、大規模言語モデルがさまざまな応用問題を解くため

に活用されている。一方、本研究では文章ではなく文書集合を対象とする。3.1 節に述べた例を当てはめると、文章の処理は「演奏」を扱い、文書集合の処理は「作曲の繰り返し」を扱うことに相当する。この違いがあるため、大規模言語モデルを本研究にそのまま当てはめることはできないが、学習させるデータを工夫するなど、何らかの工夫により深層学習の手法を本研究に適用することが考えられる。

9.3.2 再構成キーワードおよび通時的対象を示す単語への分散表現の適用

4.5 節に述べたトピック遷移構造の再構成と提示、7.2 節に延べた通時的対象の抽出では、それぞれ着目する対象が、1 つの単語として文書集合中で表されていることを前提にした。一方、例えば新たな人工物の設計過程では、ある対象の名称が後からつけられることや、他の部分と切り離して新たな対象として認識されることなどが考えられ、文書集合中で同じ対象が 1 つの単語で表現されているとは限らない。これに対応するためには、対象の表現範囲を、特定の 1 つの単語に限らず、ある程度の範囲をもって表現する方法が考えられる。

方法のひとつとして、7.2.2 節でも用いた word2vec による単語の分散表現などにより類似した単語を複数選択し、それらについて同時に処理を行う方法が考えられる。また、同様に分散表現を用い、システムの利用者が分散表現上の範囲を指定するユーザインタフェースを設けることにより、単語ではなく空間上のある範囲を対象として指定する方法も考えられる。

9.3.3 通時的対象を中心とした文書蓄積の枠組み

本論文では、扱う文書集合は所与のものとし、時間の経過に着目した処理を行った。一方、6 章で取り上げた文書集合のように、一般的な文書集合では、複数の対象について記述者が各々の記述を行っていることが多々あり、「何かが変化したということは、何か変化しないものがあることが必須である」[74] という条件に当てはまらないことが多い。

そこで、7 章にて提案した通時的対象を文書が蓄積される過程において提示することが考えられる。これにより、「変化しないもの」を文書の蓄積過程に含ませ、変化の抽出を行いやすい文書集合の蓄積を促すことができる。さらに、文書集合の蓄積を伴う本来の活動、例えば会議などの議論の場において対象を明確にすることになり、議論の進展を促すことが期待できる。

9.3.4 変化へ多様性を組み合わせることによる知的な行為の創造

本論文では文書集合内における時間経過に沿った変化に着目し、7.1 節に述べたように、変化記述型の文書集合の扱いを中心に述べた。一方、多様性記述型の文書集合も数多く存在する。同じく 7.1 節では、これらを踏まえ、「多様性記述型の文書集合から得られた情報を、通時的対象への連続した行為へと変換し、変化記述型の結果を生み出すことが、『知的な振る舞い』である

と考えられる。」と述べた。例えば、ある人工衛星の設計過程へ、これまでとは別種の新たに活用できそうな人工物を組み込む、あるいは、ある植物の成長過程に、別種の植物の成長過程を組み込む、などの形が考えられる。

このような組み合わせを実現するためには、既存の物や、物に対する手順において、相互に入れ替え可能な部分や接続可能な部分を特定する、あるいは入れ替え・接続が可能なように変形を行う必要がある。単純に自然言語上の表記であれば、大規模言語モデルを用いるなどの形でそれらしい表現を生成することが出来るが、矛盾なく実現可能な手法を自動的に発見することは、現時点では難しい。

しかしながら、それぞれから一部を切り取り、それらを接続することが新しい知識を作る [55] 方法のひとつとなることは間違いない。このために、人間と計算機が協調し、例えば基本的なアイデアと最終的な調整といった多様性をもたらす部分を人間が主に担当し、それらの変化の可能性の組み合わせの探索を計算機が行う、といった方式が考えられる。

参考文献

[1] Wikipedia entity vectors. https://github.com/singletongue/WikiEntVec. Accessed: 2021/3/24.

[2] Wikipedia:日本語版の統計. https://ja.wikipedia.org/wiki/Wikipedia:日本語版の統計. Accessed: 2021/8/2.

[3] パケット通信 (アマチュア無線)：フリー百科事典『ウィキペディア（wikipedia)』. https://ja.wikipedia.org/wiki/パケット通信_(アマチュア無線). Accessed: 2021/11/3.

[4] 中央教育審議会 議事要旨・議事録・配布資料（web archiving project）. https://warp.ndl.go.jp/info:ndljp/pid/11293659/www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/giji_list/index.htm. Accessed: 2021/3/26.

[5] Wolfgang Aigner, Silvia Miksch, Heidrun Schumann, and Christian Tominski. *Visualization of time-oriented data*. Springer Science & Business Media, 2011.

[6] J. Allan. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, 2002.

[7] David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120, 2006.

[8] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, Vol. 3, No. Jan, pp. 993–1022, 2003.

[9] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 5th edition, 2015.

[10] Samizdat Drafting Co. Arbor.js a graph visualization library using web workers and jquery. http://arborjs.org/, 2011. Accessed: 2021/2/7.

[11] Jeff Conklin and Michael L. Begeman. gibis: a hypertext tool for exploratory policy discussion. In *Proceedings of the 1988 ACM conference on Computer-supported cooperative work*, 1988.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[13] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 439–453, 07 2020.

[14] Adji B Dieng, Francisco JR Ruiz, and David M Blei. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*, 2019.

[15] John Ellson, Emden Gansner, Lefteris Koutsofios, Stephen C North, and Gordon Woodhull. Graphviz—open source graph drawing tools. In *Graph Drawing*, pp. 483–484. Springer, 2002.

[16] Paolo Federico, Wolfgang Aigner, Silvia Miksch, Florian Windhager, and Lukas Zenk. A visual analytics approach to dynamic social networks. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, pp. 1–8, 2011.

[17] David Frey, Rahul Gupta, Vikas Khandelwal, Victor Lavrenko, Anton Leuski, and James Allan. Monitoring the news: a tdt demonstration system. In *Proceedings of the first international conference on Human language technology research*, 2001.

[18] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, Vol. 21, No. 11, pp. 1129–1164, 1991.

[19] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, Vol. 46, No. 4, pp. 1–37, 2014.

[20] Google. Google ngram viewer. https://books.google.com/ngrams. Accessed: 2021/8/9.

[21] Google. Google trends. https://www.google.co.jp/trends/. Accessed: 2021/7/25.

[22] S. Havre, B. Hetzler, and L. Nowell. Themeriver: Visualizing theme changes over time. In *Proc. of IEEE Symposium on Information Visualization*, 2000.

[23] M. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, Vol. 23, pp. 33–64, 1997.

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, Vol. 9, No. 8, pp. 1735–1780, 1997.

[25] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 50–57, 1999.

[26] Patrick Jähnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. Scalable generalized dynamic topic models. In *International Conference on Artificial Intelligence and Statistics*, pp. 1427–1435. PMLR, 2018.

[27] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir tech-

niques. *ACM Transactions on Information Systems*, Vol. 20, No. 4, pp. 422–446, 2002.

[28] Y. Kitamura, M. Ikeda, and R. Mizoguchi. A causal time ontology for qualitative reasoning. In *Proc. of the Fifteenth International Joint Conference on Artificial Intelligence*, pp. 501–506, 1997.

[29] Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. https://taku910.github.io/mecab/, 2011. Accessed: 2021/2/7.

[30] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*, 2017.

[31] Rita Marques, Miguel Mira da Silva, and Diogo R Ferreira. Assessing agile software development processes with process mining: A case study. In *2018 IEEE 20th Conference on Business Informatics (CBI)*, Vol. 1, pp. 109–118. IEEE, 2018.

[32] Eva Mayr and Florian Windhager. Once upon a spacetime: Visual storytelling in cognitive and geotemporal information spaces. *ISPRS International Journal of Geo-Information*, Vol. 7, No. 3, p. 96, 2018.

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient eestimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations 2013*, 2013.

[34] Setsuo Ohsuga. Multi-strata scheme for conceptual. *Information Modelling and Knowledge Bases VIII*, Vol. 8, pp. 11–29, 1997.

[35] Vladimir Rubin, Christian W Günther, Wil MP Van Der Aalst, Ekkart Kindler, Boudewijn F Van Dongen, and Wilhelm Schäfer. Process mining framework for software processes. In *International conference on software process*, pp. 169–181. Springer, 2007.

[36] NextPhase Selling. Advanced google searching for social media. https://smperformance.wordpress.com/2011/11/09/advanced-google-searching-for-social-media/, 2011. Accessed: 2021/7/25.

[37] Simon J. Buckingham Shum, Albert M. Selvin, Maarten Sierhuis, Jeffrey Conklin, Charles B. Haley, and Bashar Nuseibeh. Hypermedia support for argumentation-based rationale –15 years on from gibis and qoc. *Technical Report KMI-05-18*, 2005.

[38] Katsuaki Tanaka and Koichi Hori. Finding diachronic objects of drifting descriptions by similar mentions. In *Pacific Rim Knowledge Acquisition Workshop*, pp. 32–43. Springer, 2019.

[39] Federico Tomasi, Praveen Chandar, Gal Levy-Fix, Mounia Lalmas-Roelleke, and Zhenwen Dai. Stochastic variational inference for dynamic correlated topic models. In *36th*

Conference on Uncertainty in Artificial Intelligence, pp. 859–868. PMLR, 2020.

[40] Alexey Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, Vol. 106, No. 2, p. 58, 2004.

[41] Wil Van Der Aalst, Arya Adriansyah, Ana Karla Alves De Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, Jagadeesh Chandra Bose, Peter Van Den Brand, Ronald Brandtjen, Joos Buijs, et al. Process mining manifesto. In *International conference on business process management*, pp. 169–194. Springer, 2011.

[42] Wil Van Der Aalst, Arya Adriansyah, Ana Karla Alves De Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, Jagadeesh Chandra Bose, Peter Van Den Brand, Ronald Brandtjen, Joos Buijs, et al. プロセスマイニングマニフェスト（最終版）. https://www.tf-pm.org/upload/1580738062276.pdf, 2012. Accessed: 2021/7/25.

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, Vol. 9, No. 11, 2008.

[44] Franz Wanner, Andreas Stoffel, Dominik Jäckle, Bum Chul Kwon, Andreas Weiler, and Daniel A. Keim. State-of-the-art report of visual analysis for event detection in text data streams. In R. Borgo, R. Maciejewski, and I. Viola, editors, *EuroVis - STARs*, pp. 125–139, Swansea, UK, 2014. Eurographics Association.

[45] Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine learning*, Vol. 23, No. 1, pp. 69–101, 1996.

[46] ダッソー・システムズ. Catia. http://www.3ds.com/jp/products/catia/. Accessed: 2021/8/2.

[47] ダッソー・システムズ. Catia - product knowledge template 1 (kt1). https://www.3ds.com/ja/products-services/catia/products/v5/portfolio/domain/Product_Synthesis/product/KT1/. Accessed: 2021/8/2.

[48] エドムントフッサール, 立松弘孝訳. 内的時間意識の現象学. みすず書房, 1967.

[49] 川島レイ. 内的時間意識の現象学 〜超小型手作り衛星、宇宙へ〜. エクスナレッジ, 2005.

[50] 長谷川幹根, 石川佳治. T-scroll: 時系列文書のクラスタリングに基づくトレンド可視化システム. 情報処理学会論文誌, Vol. 48, pp. 61–78, 2007.

[51] 環境省. 環境省中央環境審議会地球環境部会. https://www.env.go.jp/council/06earth/yoshi06.html. Accessed: 2021/3/24.

[52] 環境省. 地球環境部会について. https://www.env.go.jp/council/06earth/gaiyo06.html. Accessed: 2021/3/24.

[53] 吉川弘之, 田浦俊春, 小山照夫, 伊藤公俊. 技術知の位相. 東京大学出版会, 1997.

[54] 吉川弘之, 田浦俊春, 小山照夫, 伊藤公俊. 技術知の本質 – 文脈性と創造性. 東京大学出版会, 1997.

[55] 堀浩一. 創造活動支援の理論と応用. オーム社, 2007.

[56] 田中克明. Twitter におけるトピック遷移分析システムの提案. インタラクティブ情報アクセスと可視化マイニング研究会, Vol. 7, No. 5, pp. 22–27, jun 2014.

[57] 高藤淳, 來村徳信, 溝口理一郎. オントロジー工学に基づく技術知識統合管理システムの発展とビジネス展開. 人工知能学会論文誌, Vol. 26, pp. 547–558, 2011.

[58] 中須賀真一. 超小型衛星の研究開発. 精密工学会誌, Vol. 77, No. 1, pp. 37–41, 2011.

[59] 渡邉真也, 湊亮二郎. 多数非劣解集合からの設計支援手法の開発 – ジェットエンジン最適化を通して. 人工知能学会論文誌, Vol. 24, pp. 1–12, 2009.

[60] 人工知能学会編集委員会. 「人工知能」の表紙に対する意見や議論に関して. https://www.ai-gakkai.or.jp/whats-new/jsai-article-cover/. Accessed: 2021/9/5.

[61] 野間口大, 藤田喜久雄. 設計プロセスにおける仮説生成検証の動的展開に着目した設計支援フレームワーク. 人工知能学会論文誌, Vol. 25, pp. 514–529, 2010.

[62] 田中克明. 共起語の類似度と時刻分布を利用した文書集合からの変化記述の対象抽出の試み. 人工知能学会第 33 回全国大会論文集, p. 1N3J901, 2019.

[63] 田中勝人. 現代時系列分析. 岩波書店, 2006.

[64] 内閣府. 行政文書ファイル等の保有数 公文書等の管理等の状況 公文書管理制度 - 内閣府. https://www8.cao.go.jp/chosei/koubun/houkoku/hoyusu/hoyusu.html. Accessed: 2021/8/2.

[65] 飯島正, 田端啓一, 斎藤忍. プロセスマイニング・サーベイ (第 01 回: 概要と基本概念). 情報システム学会誌, Vol. 11, No. 2, pp. 20–53, 2016.

[66] 鳥海不二夫, 榊剛史, 岡崎直観. 「人工知能」の表紙に関する tweet の分析 (小特集「人工知能」表紙問題における議論と論点の整理). 人工知能 : 人工知能学会誌 : journal of the Japanese Society for Artificial Intelligence, Vol. 29, No. 2, pp. 172–181, mar 2014.

[67] 文部科学省. 中央教育審議会 諮問・答申等一覧. https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/toushin/index.html. Accessed: 2021/9/8.

[68] 文部科学省. 文部科学省中央教育審議会. https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo0/. Accessed: 2021/2/7.

[69] 野間口大, 高見真史, 阪口杏奈, 藤田喜久雄. 産業界における設計工学関連の手法とツールの活用状況の調査研究 (12 年間の推移に関する分析を論点として). 設計工学, Vol. 55, No. 1, pp. 43–60, 2020.

[70] 矢入健久. 機械学習とシステム同定: 動的システム学習研究の動向. 計測と制御, Vol. 58, No. 3, pp. 176–181, 2019.

[71] 本村陽一, 西田佳史. 計算論的日常生活行動理解研究基盤. 人工知能学会論文誌, Vol. 24, No. 2, pp. 284–294, 2009.

[72] 本村陽一, 西田佳史. サービス可能知識としての日常生活行動の計算モデル. 人工知能学会

論文誌, Vol. 25, pp. 651–661, 2010.

[73] 溝口理一郎. オントロジー工学. オーム社, 2005.

[74] 溝口理一郎. オントロジー工学の理論と実践. オーム社, 2012.

発表文献リスト

本研究に関連のある発表文献などを以下にあげる。

査読付き学術雑誌論文

1. Katsuaki Tanaka, Koichi Hori, Masato Yamamoto, “Development of a Recommender System based on Extending Contexts of Content and Personal History”, Journal of Emerging Technologies in Web Intelligence, Vol.2, No.3, 2010
2. Kosuke Numa, Katsuaki Tanaka, Mina Akaishi and Koichi Hori, “Reuse and Remix: Content Recomposition System based on Automatic Draft Generation”, Journal of Emerging Technologies in Web Intelligence, Vol.2, No.3, 2010
3. Kaira Sekiguchi, Katsuaki Tanaka, Koichi Hori, “‘Design with discourse’ to design from the ‘ethics level’”, Information Modeling and Knowledge Bases XXI, Frontier in Artificial Intelligence and Application, IOS Press, Vol.206, 2010
4. 田中克明, 堀浩一, 山本真人, “個人行動履歴に基づく情報推薦システムの開発”, 人工知能学会論文誌, Vol.23, No.6, 2008
5. 横山美和, 田中克明, 赤石美奈, 堀浩一, “ラジオ番組制作におけるコンテンツ生成支援に関する一手法：制作者の内省的思考を促す創造支援システム”, 映像情報メディア学会誌, Vol.59, No.5, 2005

査読付き国際会議論文

6. Katsuaki Tanaka, Koichi Hori, “Finding Diachronic Objects of Drifting Descriptions by Similar Mentions”, Proc. of 2019 Pacific Rim Knowledge Acquisition Workshop, pp.32–43, 2019
7. Katsuaki Tanaka, Koichi Hori, “Extracting Tasks in Design Process Records”, Proc. of Eighth International Joint Conference on Computer Science and Software Engineering, 2011

8. Katsuaki Tanaka, Koichi Hori, Masato Yamamoto, “A Recommender System Based on Context Extending of Content and Personal History”, Proc. of the International Conference on Information and Communication Systems, 2009
9. Kaira Sekiguchi, Katsuaki Tanaka, Koichi Hori, “Design with Discourse to Design from the Ethics Level”, Proc. of the 19th European-Japanese Conference on Information Modeling and Knowledge Bases, 2009
10. Kosuke Numa, Kiyoko Toriumi, Jun Abe, Katsuaki Tanaka, Mina Akaishi, Koichi Hori, “Content Recomposition Supported by Automatic Draft Generation”, Proc. of the International Conference on Information and Communication Systems, 2009
11. Kosuke Numa, Kiyoko Toriumi, Jun Abe, Tatsuo Sugimoto, Masako Miyata, Hideki Yoshimoto, Katsuaki Tanaka, Mina Akaishi, Koichi Hori, “Practice oriented Content Co-creation Support Systems”, Proc. of the Fifth International Conference on Collaboration Technologies, 2009
12. Katsuaki Tanaka, Mina Akaishi and Koichi Hori, “Reorganize Topic Transition in Design Process Records”, Porc. of the Third International Conference on Knowledge, Information and Creativity Support Systems, 2008
13. Kosuke Numa, Hironori Tomobe, Katsuaki Tanaka, Takuichi Nishimura, Koichi Hori, Takeshi Sunaga, “A Case Study on Interactions with User Contributed Website in Public Space”, Proc. of the 7th International Workshop on Social Intelligence Design, 2008
14. Kosuke Numa, Kiyoko Toriumi, Katsuaki Tanaka, Mina Akaishi, Koichi Hori, “Participatory Workshop as a Creativity Support System”, Proc. of12th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, 2008
15. Kosuke Numa, Kenro Aihara, Hideaki Takeda, Katsuaki Tanaka, Mina Akaishi, Koichi Hori, “Generating and Recomposing Contents using Networked Sensor Information”, Proc. of the First International Workshop on Contents Creation Activity Support by Networked Sensing in conjunction with the 5th International Conference on Networked Sensing Systems, 2008
16. Kosuke Numa, Katsuaki Tanaka, Mina Akaishi, Koichi Hori, “Activating Expression Life Cycle by Automatic Draft Generation and Interactive Creation”, Proc. of International Workshop on Recommendation and Collaboration in conjunction with 2008 International ACM Conference on Intelligent User Interfaces, 2008
17. Katsuaki Tanaka, Mina Akaishi and Koichi Hori, “Topic Change Extraction and Reorganization from Problem-Solving Records”, Proc. of Software Knowledge Information

Management and Application, 2006

18. Katsuaki Tanaka, Mina Akaishi and Koichi Hori, “Semantic Structure Transition with Elapsed Time”, Proc. of the Semantic Computing Initiative in conjunction with the 14th International Conference on World Wide Web, 2005
19. Miwa Yokoyama, Katsuaki Tanaka, Mina Akaishi, Koichi Hori, “Towards a Supporting System for Amplifying Reflective Thinking of Program Creators”, Proc. of the Language Sense on Computer in conjunction with the 8th Pacific Rim International Conference on Artificial Intelligence, 2004
20. Katsuaki Tanaka, Atsuhiro Takasu, “Topic Change Extraction from Problem Solving Records”, Proc. of the 8th World Multi-Conference on Systemics, Cybernetics and Informatics, 2004
21. Katsuaki Tanaka, Yoshikiyo Kato, Sin’ichi Nakasuka and Koichi Hori, “Using Design Information to Support Model-Based Fault Diagnosis Tasks”, Proc. of 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, 2004
22. Atsuhiro Takasu, Katsuaki Tanaka, “Feature Word Tracking in Time Series Documents”, Proc. of 5th International Conference on Intelligent Data Engineering and Automated Learning, 2004
23. Katsuaki Tanaka, Setsuo Ohsuga, “Model-Based Creation of Agents and Distribution of Problem Solving”, Proc. of the Second Asia-Pacific Conference on Intelligent Agent Technology, 2001
24. Katsuaki Tanaka, Michiko Higashiyama, Setsuo Ohsuga, “Problem Decomposition and Multi-Agent System Creation for Distributed Problem Solving”, Proc. of the 12th International Symposium on Methodologies for Intelligent Systems, 2000

査読付き国内会議論文

25. 真鍋陸太郎, 水越伸, 宮田雅子, 田中克明, 溝尻 真也, 栗原大介, “参加型コミュニティ・アーカイブのデザイン：デジタル・ストーリーテリングや参加型まちづくりの融合”, デジタルアーカイブ学会誌, Vol.4, No.2, pp.113-116, 2020

招待講演

26. 田中克明, 堀浩一, “蓄積情報からの変化の抽出と再構成— 小型衛星設計と個人行動履歴を例に—”, 第 6 回知識創造支援システムシンポジウム招待講演, 2009

解説など

27. 田中克明, “『5』編集室・田中克明の一口メモ”, “特集：人と機械が生み出す新世界”, 5: Desingning Media Ecology, Vol.9, 2018
28. 田中克明, 宮田雅子, “SYSTEM0706 のひみつ”, 5: Desingning Media Ecology, Vol.2, 2014
29. 堀浩一, 赤石美奈, 田中克明, 沼晃介, “表現を生むための場としてのワークショップを支える人工知能技術”, 人工知能学会誌, Vol.26, No.5, 2011
30. 濱崎雅弘, 久保田秀和, 江渡 浩一郎, 中村嘉志, 田中克明, 西村 拓一, “表現の連鎖を支える技術”, 人工知能学会誌, Vol.26, No.5, 2011

紀要など

31. 田中克明, “共起語の類似度を利用した文書集合からの変化記述の対象抽出の試み”, 埼玉工業大学人間社会学部紀要, Vol.17, pp.15–21, 2019

査読無し国内会議

32. 田中克明, “共起語の類似度と時刻分布を利用した文書集合からの変化記述の対象抽出の試み”, 2019 年度人工知能学会全国大会予稿集, 2019
33. 田中克明, “黒電話により個人の語りを蓄積する仕組みの検討”, 人工知能学会 第 1 回市民共創知研究会予稿集, 2016
34. 田中克明, “Twitter におけるトピック遷移分析システムの提案”, インタラクティブ情報アクセスと可視化マイニング研究会第 7 回予稿集, 2014
35. 田中克明, 濱崎雅弘, “異分野共同研究履歴分析の事例”, 2013 年度人工知能学会全国大会予稿集, 2013
36. 濱崎雅弘, 沼晃介, 田中克明, “異分野越境型プロジェクトにおけるコミュニケーションとコラボレーションに関する一考察”, 2012 年度人工知能学会全国大会予稿集, 2012
37. 田中克明, 濱崎雅弘, 小早川真衣子, 堀浩一, “オフライン世界とオンライン世界における協調的創造活動の違いの考察”, 2010 年度人工知能学会全国大会予稿集, 2010
38. 田中克明, 堀浩一, “Twitter ハッシュタグに基づく Tweet 群からの変化抽出”, 電子情報通信学会 WI2-2010-1〜13, 2010
39. 田中克明, 沼晃介, 堀浩一, “創造活動における表現変化の抽出と利用の検討”, 2009 年度人工知能学会全国大会予稿集, 2009
40. 佐藤一夫, 山本真人, 小林功, 佐治信之, 田中克明, “行動履歴に基づく情報推薦基盤と推論

エンジンの開発”, 電子情報通信学会技術研究報告人工知能と知識処理, Vol.108, No.119, 2009

41. 田中克明, 堀浩一, 山本真人, “表現の他者文脈への伸延による流通促進の試み”, 2008 年度人工知能学会全国大会予稿集, 2008
42. 沼晃介, 田中克明, 赤石美奈, 堀浩一, “表現候補の自動生成とインタラクションに基づく表現の液状化・結晶化サイクルの促進”, 2008 年度人工知能学会全国大会予稿集, 2008
43. 佐藤真, 田中克明, 赤石美奈, 堀浩一, “物語構造モデルに基づき話題の遷移を分析する手法の提案”, 2008 年度人工知能学会全国大会予稿集, 2008
44. 田中克明, 赤石美奈, 堀浩一, “設計議事録からの主題遷移構造の抽出と利用”, 2007 年度人工知能学会全国大会予稿集, 2007
45. 唐澤悠紀, 田中克明, 赤石美奈, 堀浩一, “ドキュメント群を利用した設計プロセス支援”, 2007 年度人工知能学会全国大会予稿集, 2007
46. 佐藤真, 田中克明, 赤石美奈, 堀浩一, “視覚情報から抽出した文脈を用いた情報アクセス・システムの提案”, 2007 年度人工知能学会全国大会予稿集, 2007
47. 石川敏照, 田中克明, 赤石美奈, 堀浩一, “失敗例と事故例からの事故予測”, 2007 年度人工知能学会全国大会予稿集, 2007
48. 関口海良, 田中克明, 赤石美奈, 堀浩一, “アクティブなヒューマノイドのネットワーク化の提案—動的な哲学知の生成と多元主義の実践を目指して—”, 2007 年度人工知能学会全国大会予稿集, 2007
49. 佐藤真, 田中克明, 赤石美奈, 堀浩一, “物語構造モデルに基づく話題類似連鎖抽出”, 電子情報通信学会技術研究報告知能ソフトウェア工学, Vol.107, No.159, 2007
50. 田中克明, 赤石美奈, 堀浩一, “設計議事録からの主題階層構造変化の抽出”, 人工知能学会第 65 回人工知能基本問題研究会予稿集, 2007
51. 田中克明, 赤石美奈, 堀浩一, “設計議事録からの設計プロセス抽出の試み”, 電子情報通信学会技術研究報告知能ソフトウェア工学, Vol.106, No.472, 2007
52. 関口海良, 田中克明, 赤石美奈, 堀浩一, “ロボットは会議に潜む多重文脈の表出を支援できるか”, 2006 年度情報処理学会全国大会講演論文集, 2006
53. 田中克明, 堀浩一, “対象変化の再構成による設計支援”, 第 7 回 AI 若手の集い予稿集, 2006
54. 田中克明, 赤石美奈, 高須淳宏, 堀浩一, “設計議事録からの主題構造変化の抽出と再構成”, 人工知能学会第 59 回人工知能基本問題研究会予稿集, 2005
55. 横山美和, 田中克明, 赤石美奈, 堀浩一, “番組制作者の内省的思考を促す創造支援システム”, 人工知能学会第 18 回ことば工学研究会予稿集, 2004
56. 田中克明, 加藤義清, 大須賀節雄, 堀浩一, “大規模問題における問題構造に応じた知識の収集”, 電子情報通信学会技術研究報告知能ソフトウェア工学, Vol.103, No.217, 2003

57. 田中克明, 複数システム統合による衛星故障診断, 第 4 回複雑システムの科学技術シンポジウム, 2003
58. 田中克明, 大須賀節雄, “設計者を取り込んだ設計支援システムの試み”, 2002 年度人工知能学会全国大会予稿集, 2002

口頭発表など

59. Shin Mizukoshi, Rikutaro Manabe, Katsuaki Tanaka, “Digital Storytelling”, Doing Digital Methods: Interdisciplinary Interventions, 2018
60. 水越伸, 宮田雅子, 真鍋陸太郎, 田中克明, 栗原大介, “テレフォノスコープ：電話機型装置によるマイクロ・デジタル・ストーリーテリング”, カルチュラル・タイフーン 2016, 2016
61. 田中克明, “ネビュラに向けていくつかの視角から”, MELL Platz 第 19 回公開研究会, 2010

謝辞

本論文をまとめるにあたり、たくさんの方々にご指導、ご助力を賜りました。

本論文の主査である堀浩一教授には、本当に長い期間に渡り、さまざまなご指導をいただきました。「知識は速度のようなものであり、観測はできるがしまっておくことはできない」という堀先生のお話が、人工知能研究における本研究の大きな目標となり、ここまで研究を進めることができました。私が現在もこうして研究を進められているのは、先生の存在があってこそであり、心から厚く感謝申し上げます。

東京大学大学院工学系研究科航空宇宙専攻の中須賀真一教授、同じく岩崎晃教授、東京大学先端科学技術研究センターの矢入健久教授、法政大学情報科学部コンピュータ科学科の赤石美奈教授には、本研究への的確かつ有益なご助言を、数多く賜りました。とくに、中須賀先生からは、本研究で中心的に扱った CubeSat XI-IV の設計議事録をご提供いただいくことで、本研究の原動力とすることができました。また、赤石先生からは、本研究の初期より数多くのご助言をいただきました。トピック遷移構造の再構成は、赤石先生のご研究から大きな影響を受けた部分でもあります。あらためて、深く御礼申し上げます。

国立情報学研究所の高須淳宏教授には、本研究を開始する大きなきっかけをいただきました。文書クラスタリングを時間方向に繰り返すアイデアを研究として進めることができたのは、高須先生のお力添えをいただいたことが大きく、深く感謝いたします。

東京大学名誉教授の大須賀節雄先生には、私が学部学生であった当初から、非常にたくさんのご指導をいただきました。深く感謝の意を表します。大須賀先生の設計の自動化、多階層モデルの研究は私の研究の出発点であり、多階層モデルを時間方向に拡張することが、いまでも私の研究の目標です。

もう 20 年も経ってしまいましたが、早稲田大学大学院理工学研究科情報科学専攻（当時）の大須賀研究室のメンバーにも、感謝を述べたいと思います。同じく、あっという間に 10 年以上が経過しましたが、東京大学工学系研究科航空宇宙工学専攻堀研究室のみなさまにも、深く感謝の意を表します。

東京大学大学院情報学環の水越伸教授には、研究のアイデアを実践する機会を、数多くいただきました。本論文の一部として直接まとめることはかないませんでしたが、私が持てる研究のア

イデアを実践する機会を持ち続けることができたことは、水越先生のお力によるところが大きく、深く御礼申し上げます。

一橋大学情報基盤センターのみなさま、埼玉工業大学人間社会学部情報社会学科のみなさま、株式会社多聞のみなさまにも、さまざまな形でお力添えをいただきました。深く感謝の意を表します。また、埼玉工業大学人間社会学部情報社会学科知能情報システム研究室のメンバーには、本論文の実験に協力いただきました。深く感謝します。

人工知能研究が、人の活動の表層を写すのではなく人の活動と同じところに至るには、時間を積み重ねた思索の要素を取り入れる必要があると考え、まだその端緒に過ぎないこの論文をまとめるのに、ずいぶん時間がかかってしまいました。その期間に非常に多くの方にお力添えいただきました。個々にお名前を挙げることはできませんが、深く感謝を申し上げます。

最後に、本論文のシステムの評価、さらに論文の執筆にあいて、UWC Adriatic の川崎健開氏、株式会社多聞の平林園絵氏には、なみなみならぬご助力をいただきました。また、田中理名氏からは、いつも仕事の手伝いを申し出ていただきました。あらためて、感謝を表します。