様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２４年６月１３日現在

機関番号：１２６１３
研究種目：若手研究(B)
研究期間：2008～2011
課題番号：２０７００１２９
研究課題名（和文）　時系列文書における主題遷移パターンの抽出と利用

研究課題名（英文）　Extraction and reorganization of topic transition patterns in recorded document

研究代表者
田中　克明（TANAKA KATSUAKI）
一橋大学・情報基盤センター・助教
研究者番号：80376657

研究成果の概要（和文）：本研究では、時間経過とともに蓄積された文書群が含む「時間の経過」に着目し、文書群の中に記述された主題とその変化の抽出と、文書群利用者の目的に合わせた再構成を行うシステムを提案した。文書群をより細かい断片に分けこれらをクラスタリングすることにより、時間経過の中のある時点での主題群を抽出することを基本とし、時間の経過とともに新たな文書の追加や古い話題を構成する断片の忘却を行うことで、時間経過に沿った主題の遷移を抽出した。また、主題について、一時的なもの、定常的に出現するものに分類するとともに、これらを指定したキーワードをもとに再構成し、提示するシステムを構築した。

研究成果の概要（英文）：This research proposed a system which focused chronological order of documents had accumulated over a period of time. It extracted and reorganized topics and its transitions from documents to acquire knowledge.　It decomposed each document into smaller text fragments and grouped them into clusters of fragments.　We regarded a cluster of text fragments as a topic.　The system made document groups based on elapsed time, and calculated relevance of topics in neighboring document groups. Then it obtained topics and their transition. It could sort out topics temporal topics and constant topics. It also reorganized topics and their transition with keywords indicated by a user.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2008 年度	800,000	240,000	1,040,000
2009 年度	900,000	270,000	1,170,000
2010 年度	900,000	270,000	1,170,000
2011 年度	600,000	180,000	780,000
総　計	3,200,000	960,000	4,160,000

研究分野：情報学
科研費の分科・細目：情報学・知能情報学
キーワード：情報システム、人工知能、可視化、設計工学

１．研究開始当初の背景

概念・知識を体系化する枠組としてオントロジーが提案された初期には、文書から計算機を用いてオントロジーを自動構築する試みがいくつかなされた。しかし、これらは広

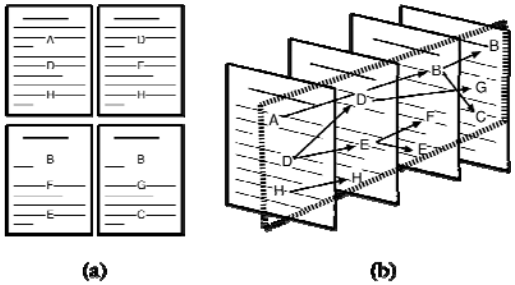

図 1: 文書群の時間経過を意識しない扱い(a)と時間経過を意識した扱い(b)

く用いられるに至っていない。オントロジーが対象とする知識は、すでにある程度固定されたものであるのに対し、計算機によるオントロジー構築において知識源として用いられたのは、さまざまな人が記述した文書群であり、その内容は時間の経過とともに変化するものであった。

このように、動的に変化する知識源を扱いながら、それを静的な変化しないものへと変換しようとしたことが、計算機によるオントロジー自動構築が成功しなかった要因のひとつであると考えられる。すなわち、文書は図 1(b)に示すように時間経過に沿った内容の変化を含むものであるが、これに留意せず、各文書内部に記述された内容を抽出(図 1(a))してオントロジーに包含させる知識としたことが理由のひとつとして考えられる。

そこで、時間経過にともなう文書の記述内容の変化に着目し、本研究を開始した。とくに、時間経過に沿って観察される記述内容の変化について、変化が知識の利用によって引き起こされるものであると考え、時間経過に沿って蓄積された文書から変化を取り出すことができれば知識の発見・体系化の手がかりとなるとの着想から、文書群における時間経過にともなう主題変化の抽出と利用に関する研究を進めた。

２．研究の目的

本研究では、時間経過の中で使用された知識が文書の記述に変化をもたらしていると考え、比較的長期間にわたって蓄積された文書群を対象として解析を行い、記述された主題とその変化を抽出する。また、変化を 1. 変化をもたず定常的に記述される主題（オントロジー的なもの)、2. ある時点でのみ使われる主題（アドホックに使われるヒューリスティクス的なもの）といったパターンに分類すること、および抽出した変化の利用を行う。

そのために、文書からの記述内容の時間変化の抽出と、それらを人間の視点に基づいて動的に再構成することができるシステムを構築する。

本研究で用いる主題抽出手法では、処理対象の文書をより小さな断片に分解する。文書分解の際の粒度を何通りかの場合にわけ、これに基づき主題抽出計算を行い、目的に適した粒度を見つける。この粒度のサイズは、対象とする文書群に依存するため、随時計算が必要である。計算を効果的に行うため、どのような性質の文書群に対しどの程度の粒度を設定することが適切か、定性的な指針を見つけることを第一の課題とする。

また、本研究では、言及がない主題に含まれる文書断片の重みを減らすことで、主題を忘却するモデルを導入している。この忘却における重み減少の度合いその他のパラメータ設定と、一時的、あるいは定常的であると分類できる主題内容との関連を見つけることを、第二の課題とする。

第三の課題として、様々な文書群に本研究の手法を適用し、どのような文書群において本手法による知識発見の支援が有効であるかの確認を行う。

３．研究の方法

本研究ではまず、蓄積された文書群に記述されているある特定の内容に関連したまとまりを主題とみなし、主題と時間経過に沿った変化の抽出を以下の手順で行う。

1. 文書作成時刻による文書集合定義
2. 文書の断片化
3. 文書集合に以下を繰り返す
   - 3-1. Probabilistic Latent Semantics Indexing （pLSI）を用いた文書断片のクラスタリング（主題抽出）
   - 3-2. クラスタリング結果に基づく主題（クラスタ）の風化処理
4. 隣接文書集合間の主題の関連度計算による主題遷移の抽出

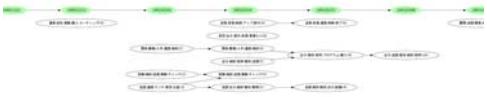

図 2: 主題とその変化の抽出例

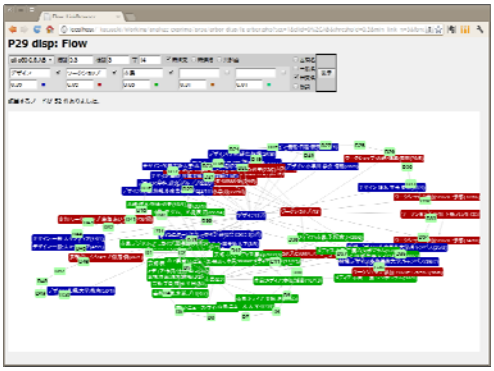

図 3: 主題とその変化の再構成例

主題とその変化の抽出に続いて、一時的な主題・定常的な主題の分類、人間の視点に合わせた主題とその変化の再構成を行う。

これらの手順の実現と、実際に計算機上で処理を行うシステムの構築を通して本研究を進めた。

4. 研究成果

(1) 文書群からの主題とその変化の抽出と再構成

まず、本研究の方法として述べた時間経過に沿って蓄積された文書群から、そこに記述された主題とその変化を抽出する手法を計算機上に実装した。実装したシステムによる主題とその変化の抽出例を図 2 に示す。

また、本研究では変化は知識の利用によって引き起こされるものであると考え、抽出した主題とその変化は全体を俯瞰するだけではなく、ある視点に関連した知識が用いられた形跡を確認しやすいように主題とその変化を再構成し、表示する機能を設けた。この機能では、視点として与えた単語に関連する主題を選び、関連する主題を連結して表示する。主題とその変化を再構成し表示した例を図 3 に示す。

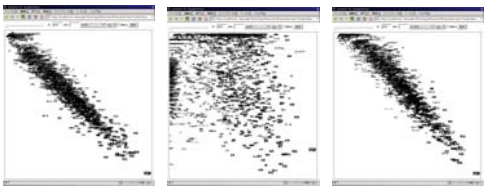

図 4: 断片長ごとの pLSI による出現単語の 2 次元配置（左から断片長 400、800、1600 バイト）

(2) 文書断片化の長さと主題抽出の関連

一定上の長さを持つ文書には、複数の内容が記述されている。本研究で用いる pLSI や近年多く利用されている LDA など潜在意味分析による文書分析手法では、文書が複数の主題を含むことを前提としてモデルが構築されている。しかし、例えば会議の議事録のように、ひとつの「文書」でも議題ごとに記述内容が大きく違う場合、文書を小さな断片に分割したうえで分析を行う方が、主題の抽出精度が向上すると考えられる。そこで、文書を簡易な手法で分割、断片化し、潜在意味分析にどのような影響を与えるかを調べた。例として、小型人工衛星の設計議事録に対し、一定の長さで断片化を行うという設定の下、断片長を変化させ pLSI を適用し、第 1 主成分、第 2 主成分をそれぞれ x 軸、y 軸として、出現単語の分布を描画した結果を図 4 に示す。この例では、断片化を行う際の断片長を 800 バイトとした場合に、pLSI の第 1、第 2 主成分に着目した際の分類が効果的に行えることがわかる。

この他、メーリングリストのメール、Twitter から取得したツイートなどを対象に処理を行い比較した結果、議事録などの「文書」の方が断片長をある程度まで長くすると次数の小さい主成分の付与率が上がる傾向を示すものの、絶対的にどの程度の長さが良いかは、文書に依存する事がわかった。

(3) 定常的な主題と一時的な主題の分類

定常的な主題と一時的な主題を分類するために、文書群ではなく、実際に人が日常行っている行動の履歴に対して主題と変化の抽出手法を用い、個人が日常的な行動をとっているのか、非日常的な行動をとっているのかを判別する仕組みの研究を行った。個人の行動履歴の取得には携帯電話を利用し、これ

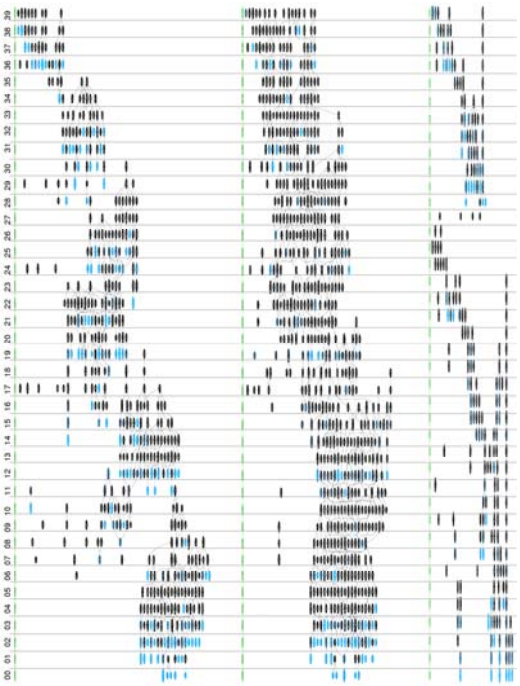

図 5: 行動履歴のパラメータ調節を伴う分類例
（水色が一時的なもの、黒が定常的なもの）

を通して閲覧や記述したコンテンツと位置の履歴を蓄積して解析対象とした。

蓄積文書群に対するシステムを応用し、行動履歴から主題として行動履歴の集合を抽出し、新しい履歴を含まない主題を忘却するものとして処理対象から徐々に外すことにより、忘却されずに残り続ける日常的な行動と、非連続に現れる非日常的な行動とを区別して提示を行うシステムを構築した。

このシステムを用い、蓄積された個人の行動履歴を、いくつの主題に分類するか、忘却を行わせる際の忘却度合いをどの程度にするか、などのパラメータ調節を行うことにより、「日常の行動」と「非日常の行動」の分類を行うことができた（図 5）。

さらにシステムを時間経過に沿って蓄積された文書群に適用し、「日常の行動」が定常的な主題に、「非日常の行動」が一時的な主題に相当する形で同様に抽出できることを確認した。

(4) 主題・単語分布のアニメーション表示

抽出された主題について、pLSI における第 1、第 2 主成分に着目し、各時刻において出現している主題の分布を平面上に描画し、

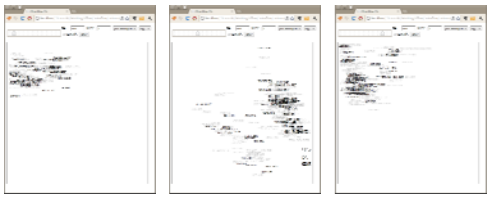

図 6: 時間経過に沿った主題分布のアニメーション表示（左から文書群の初期、中期、終期）

時間経過にそって変化させ、アニメーションとして提示する仕組みを設けた（図 6）。このアニメーション表示では、定常的な主題は同じ位置に一定期間出現したままとなる一方、一時的な主題は描画されて消えるため、主題が定常的なものか、一時的なものか、人間の直感にそって把握を行うことが可能であった。

(5) 異なった種類の文書への適用

近年 Web 上に数多く蓄積されるようになった、多数の人間が参加するコミュニケーションの記録を対象として、主題とその遷移の抽出を試みた。時系列に沿って蓄積される文書から主題とその遷移を取り出すという本研究の基本的なモデルは変更せず、文書から主題を抽出する文書の断片化とクラスタリングの手法、忘却のためのパラメータ設定などを検討した。また、抽出した主題とその変化の再構成をどのような視点に基づいて行うかは、利用者側の背景知識に依存するため、これを刺激することを目的とし、再構成のキーとなる単語を計算機システム側から提案する仕組みを取り入れた。

これらの結果、Web 上のコミュニケーションにともない蓄積された文書群から利用者が再構成を行いやすい主題の遷移の抽出を行い、利用者が再構成を行うことで何らかの知見を得やすいかには、主題抽出の手法、パラメータの設定とあわせて、対象とする文書群の種類や選択の方法が、大きな影響を持つことが分かった。

たとえば、文書が「人工衛星」のような特定の対象も「人工衛星を設計する」といった共通の目的も持たず、たとえば相互に情報交換を行うように、多数の対象について並行して議論を行っている場合、議論にまつわる複数の時間が並行して進んでいる状態となり、

一律に忘却を行う現状のモデルでは、主題遷移の抽出、および再構成の結果が、解析者にとって理解しづらいものとなる。

これより、本研究が用いる手法は、文書の作成者たちが自分自身の考える対象をそれぞれに持ち記述を行った文書群よりも、設計記録のように、文書の作成者たちが共通の対象に関わり対象を成長させている過程を記録した文書群に対して、効果的であるということが分かった。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 5 件）

① Katsuaki Tanaka, “Extracting Tasks in Design Process Records”, Proc. of the 8th International Joint Conference on Computer Science and Software Engineering, pp.373-378, 2011, 査読有
② 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, pp.197-203, 2010, 査読有
③ Kosuke Numa, Katsuaki Tanaka, Mina Akaishi, Koichi Hori, “Reuse and Remix: Content Recomposition System based on Automatic Draft Generation”, Journal of Emerging Technologies in Web Intelligence, Vol.2, No.3, pp.191-196, 2010, 査読有
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, pp.457-464, 2008, 査読有
⑤ Katsuaki Tanaka, Mina Akaishi, Koichi Hori, “Reorganizing Topic Transitions in Design Process Records”, Proc. of the Third International Conference on Knowledge, Information and Creativity Support Systems, pp.148-155, 2008, 査読有

〔学会発表〕（計 5 件）

① 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”, 第 24 回人工知能学会全国大会, 2010 年 6 月 10 日, 長崎
② 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 Web インテリジェンスとインタラクション研究会, 2010 年 3 月 15 日, 大阪
③ 田中克明, 堀浩一, “創造活動における表現変化の抽出と利用の検討”, 第 23 回人工知能学会全国大会, 2009 年 6 月 19 日, 高松
④ 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 年 2 月 27 日, 石川県
⑤ 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”, 電子情報通信学会人工知能と知識処理研究会, 2009 年 1 月 22 日, 東京

６．研究組織

(1) 研究代表者

田中 克明（TANAKA KATSUAKI）
一橋大学・情報基盤センター・助教
研究者番号：80376657