3C2-3

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

Topic Transition Analysis based on Narrtivity Structure Model

佐藤 真[*1] 田中 克明[*2] 赤石 美奈[*2] 堀 浩一[*2]
Makoto Sato Katsuaki Tanaka Mina Akaishi Koichi Hori

[*1]東京大学大学院工学系研究科航空宇宙工学専攻
Department of Aeronautics and Astoronautics, School of Engineering, University of Tokyo

[*2]東京大学先端科学技術研究センター
Research Center for Advanced Science and Technology, University of Tokyo

This paper proposes a method that compose a new story from scenes. We define context-dependent term attractiveness with Narrativity Structure Model based on the idea that composition of scenes determines the weight of terms that each scene contains. We then show examples of its implementation to extract topic transitions from design records.

1. はじめに

設計過程において新しい問題が発生したとき，以前の問題に対する解決方法をそのまま利用できることは少ない．すなわち新しい問題を解決するためにはそれまでに蓄えてきた様々な知識を組み合わせ新しい知識を創り出し解決方法を編み出す必要がある．私たちはこのようなプロセスを計算機によって支援するためには情報の分解・再構成のフレームワークを開発することが有効であると考えている．

本研究では分解・再構成のフレームワークの開発の一端として，文書断片の構成の仕方によって各々の文書断片の意味が変化するのでないかという考えに基づき，文書断片をつなぎ合わせ再構成するための手法を提案する．そして提案手法を実装したシステムの利用例を東京大学中須賀研究室の小型人工衛星の設計議事録を対象に示す．

2. ナラティブ連想情報アクセス

本章では情報の分解・再構成のフレームワークとしてナラティブ連想情報アクセス [赤石 06] について述べる．

ナラティブ連想情報アクセスは物語構造モデルに基づき情報にアクセスするフレームワークである．これは大量の文書を対象に情報が必要とされている文脈に応じて動的に情報を分解・再構成を行い潜在的な物語を紡ぎだすものである．物語構造モデルは知識を伝達するためには物語性が重要であるという認識と大量の文書に対して分節を行うという目的に基づき，物語言説から得られる表層的な特徴量を用いて物語内容の構造を抽出するためのモデルである．文書の構成要素に対しては表 1 のように対応づける．

ナラティブ連想情報アクセスを可能とする Narrative Navigator (NANA)（図 1）は，文書を分解する decomposition unit と再構成する composition unit からなる．語と語の連鎖関係を語の共起依存度と吸引力により語彙連鎖グラフに変換しグラフの構造に対する操作を通じて文書の分解・再構成を行う．

連絡先: 佐藤 真，東京大学大学院工学系研究科航空宇宙工学専攻，〒 153-8904 東京大学先端科学技術研究センター 4 号館 513 号室，(03)5452-5289，satomakoto[at]ai.rcast.u-tokyo.ac.jp

表 1: 物語構造モデルと文書構成要素

world model	set of stories
story	sequence of scenes (documents)
scene	chunk of event
event	set of terms (sentence)
character	term

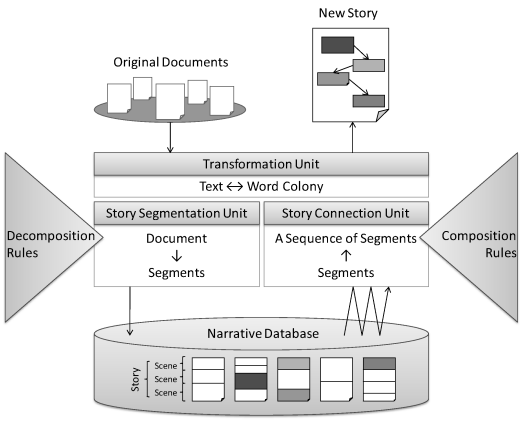

図 1: Narrative Navigator

2.1 共起依存度 (term dependency)

文書を構成する N_t 個の語 $t_i(i = 1 \sim N_t)$ について，語 t_i から語 t_j への共起依存度 $td(t_i, t_j)$ は，語 t_i が出現した同じ文中に語 t_j が出現する条件付き確率で定義する．共起依存度 $td(t_i, t_j)$ は，語 t_i が出現した文の数を $sentences(t_i)$，語 t_i と語 t_j が同時に出現した文の数を $sentences(t_i, t_j)$ とすれば，以下の式で計算される．

$$td(t_i, t_j) = \frac{sentences(t_i, t_j)}{sentences(t_i)} \quad (1)$$

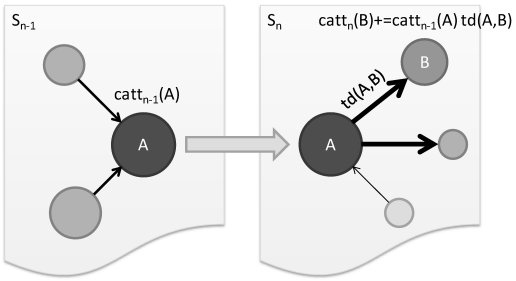

図 2: 文脈依存吸引力の概念図

2.2 吸引力 (term attractiveness)

吸引力 $attr(t_i)$ は，語 t_i が文書中の他の語を引き付ける力を，他の語から語 t_i に対する出現依存度の総和として定義する．吸引力の大きさは文書中の語の重要度を示す指標となる．以下の式で計算される．

$$attr(t_i) = \sum_k td(t_k, t_i) \tag{2}$$

2.3 トピック遷移パターン分析

トピック遷移パターン分析は NANA の composition unit のための composition rule であり，ユーザの連想を促し新しい場面をつなげさせていくための指針を導く．ある文書において吸引力の大きい語をメイン・トピック語，それ以外の語をサブ・トピック語とするとき，ユーザの連想を促すためには前場面のあるメイン・トピック語がメイン・トピック [M to M] あるいは前場面のあるサブ・トピック語がメイン・トピック [S to M] となる場面を網羅的に探索し連結する場面の候補として提示することが有効であると考える．

3. 提案する手法

文書を分解・再構成する上で文書断片の持つ意味を「何としてみなすか」という手法は必要不可欠である．文書の意味理解は読者の知識や状況による一方で文書の構成にも起因するとされている．そこで文書断片の構成の仕方によって各文書断片における語の関係や重要語が変化するのではないかと仮説を立て「文脈依存吸引力」を提案する．なお物語構造モデルに基づき適当な粒度の文書断片を「場面（Scene）」と呼ぶこととする．

3.1 文脈依存吸引力

前章で述べた共起依存度と吸引力の概念を拡張し，前の場面で重要である語に次の場面で依存される語がより吸引力を持つと考える．場面のつながりを考慮した語の重要度の指標として，文書を構成する N_t 個の単語 $t_i (i = 1, \cdots, N_t)$ の場面を n 回遷移したときの文脈依存吸引力（context-dependent term attractiveness）$catt_n(t_i)$ を遷移前の場面における語の文脈依存吸引力と遷移後の場面における語の共起依存度の積の総和と定義する．

$$catt_n(t_i) = \sum_{k=1}^{N_t} catt_{n-1}(t_k) td_{S_n}(t_k, t_i) \tag{3}$$

ここで $td_{S_n}(t_i, t_j)$ は n 回目に遷移した場面 S_n における語 t_i から語 t_j への共起依存度とする．図 2 に概念図を示す．なお初期文脈依存吸引力 $catt_0(t)$ は適当に与える必要がある．

なおこの概念は，語はノード，文脈依存吸引力はノードの活性値，共起依存度行列はリンク構造とすれば Spreading Activation Model[Anderson 83] とみなすこともできる．

3.2 ベクトル空間モデル

ベクトル空間モデル [Salton 75] に倣い，文書の表す話題を，文脈依存吸引力を成分とするベクトルとみなす．

i 番目の成分を語 t_i の文脈依存吸引力とする文脈依存吸引力ベクトル $\mathbf{c}_n$ を定義する．

$$\mathbf{c}_n = \left[\begin{array}{ccc} catt_n(t_1) & \cdots & catt_n(t_{N_t}) \end{array} \right]^T \tag{4}$$

i 行 j 列の成分を場面 S における共起依存度 $td(t_j, t_i)$ とする行列を共起依存度行列 $\mathbf{D}_S$ を定義する．

$$\mathbf{D}_S = \left[\begin{array}{ccc} td_S(t_1, t_1) & \cdots & td_S(t_{N_t}, t_1) \\ \vdots & \ddots & \vdots \\ td_S(t_1, t_{N_t}) & \cdots & td_S(t_{N_t}, t_{N_t}) \end{array} \right] \tag{5}$$

共起依存度行列と文脈依存吸引力ベクトルの間には

$$\mathbf{c}_n = \mathbf{D}_{S_n} \mathbf{c}_{n-1} \tag{6}$$

という関係が成り立つ．

文脈依存吸引力ベクトルは話題の方向性を特徴づけるものである．よって提案した手法では各場面の話題をベクトル（文脈依存吸引力ベクトル）としたとき，話題の遷移を場面における語の関係行列（共起依存度行列）による線形変換であるとみなせる．

3.3 吸引力と文脈依存吸引力の比較

吸引力と文脈依存吸引力に関して次の仮説を立てる．

- 吸引力の大きさは，場面ごとに定まる重要度の指標となる
- 文脈依存吸引力の大きさは，前の場面の内容を踏まえた上での語の重要度の指標となる

これを検証するために，文書から吸引力と文脈依存吸引力を求めその結果を比較する実験を行った．

東京大学中須賀研究室の小型人工衛星の設計議事録から適当に 3 件（A，B，C）を選び，A，B，C それぞれの吸引力と，B，C の吸引力を初期文脈依存吸引力として A に遷移したときの文脈吸引力を求めた．なお共起依存度を計算するにあたり語として名詞のみを扱った．名詞の抽出に関しては手作業によって作成したオリジナル辞書を用いた．

表 2 に各場面の吸引力上位 10 語を示し，表 3 に A の吸引力上位 10 語と B，C から A へ遷移したときの文脈依存吸引力上位 10 語を示す．仮説のとおりならば，それぞれ前の場面を踏まえて A の中でどの単語が重要であるかを示している．

これを検証するために，A の吸引力では 10 位以内に入っていなかったが，(B,A) の文脈依存吸引力において 10 位以内に入っている語に着目した．「アンテナ」「SEL」は A にも B にも出現する語であり，「場面 B で重要であった語」が A 内でさらに重要度が高まった．また「USSS」「TNC」は B には出現しないが重要度が高まっている語であり，「『B で重要であった語』に関係する語」も A において重要度が高まった．

吸引力は場面の情報を静的とみなしたときの語の重要度を示し，文脈依存吸引力は場面の情報を動的とみなしたときの語の重要度を示すと考えられるため，これら 2 つの指標は必要とされる状況によってそれぞれに有意性があると考えられる．

表 2: 各場面の吸引力の上位 10 語

	A	B	C
1	電源系	FET	太陽電池
2	アンテナ	MOS	電池
3	OBC	アンテナ	意味
4	基板	振動	姿勢
5	実験	熱	回復
6	SEL	短絡	収支
7	検知	展開	ダウンリンク
8	TNC	衛星	ON
9	アマチュア	リセット	電力
10	無線	受け	電圧

表 3: 場面 A の吸引力と文脈依存吸引力の上位 10 語

	A	(B,A)	(C,A)
1	電源系	アンテナ	電源系
2	アンテナ	電源系	電力
3	OBC	OBC	ボックス
4	基板	SEL	鵜川
5	実験	検知	作成
6	SEL	アマチュア	早急
7	検知	無線	電池
8	TNC	USSS	送信
9	アマチュア	TNC	電圧
10	無線	Poly	出力

4. 実装と適用

筆者らは設計の場において意識していたにも関わらず忘れてしまったことや意識すること自体が少ないことをあえて指摘するために，映像情報に出現するオブジェクトの出現依存関係を抽出する手法を提案し，この出現依存関係を用いて既存の情報にアクセスするシステムの概念設計を行った [佐藤 07]．この情報アクセス・システムは出現依存関係をクエリとして文書の再構成を行うというものである．すなわち出現依存関係をネットワークとみなせば，そのネットワークを骨格にノード（語）間の情報（場面）の連鎖関係を抽出することでこのネットワークをよりリッチなものにでき，さらにこのネットワークを構成するために用いた場面の情報を利用することで文書の検索にも役立つのではないかと考えた（図 3）．本研究では文書群から 2 つの語を結ぶ多重文脈における場面の連鎖構造（図 4）を生成するために提案手法を実装したシステムの構築および実験を行った．

4.1 実装

システムの機能は「2 つの語の間に潜在する，ある文脈における場面の連鎖構造を抽出する」ことである．場面の連鎖構造を生成するためには以下の手順を踏む．

ユーザは初めの語と終わりの語を指定する．システムは指定された場面に応じた連結候補場面を提示する．ユーザは候補の中から連結する場面を選ぶ．システムは選ばれた場面をつなげる．つなげたら，さらに次の場面の候補を挙げる．満足のいくまでこれらのプロセスを繰り返す．

連結候補場面を提示するために，文脈依存吸引力に基づくトピック遷移パターンを利用する．[赤石 06] では吸引力の大きい語をメイン・トピック語と用いていたが，本研究では文脈依存吸引力の大きい語を用いることとする．これにより情報断片の動的な扱いを可能にする．

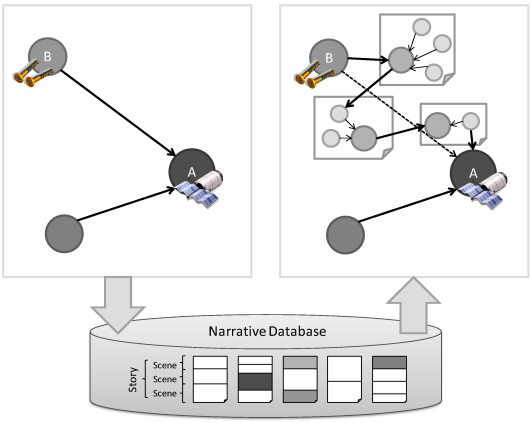

図 3: クエリとしてのネットワーク

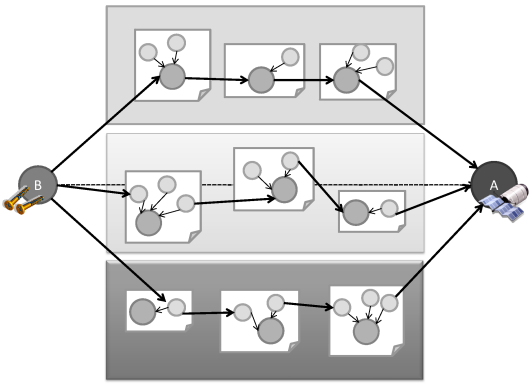

図 4: 2 つの語の間にある多重文脈

4.2 適用

2 つの語「電池」「問題」結ぶ多重文脈における場面の連鎖構造を生成するために，文脈依存吸引力によるトピック遷移パターンによる場面再構成を試みた．

東京大学中須賀研究室の小型人工衛星の設計議事録から 200 件の文書を 837 個の文書断片にしたものを対象として，システムの文脈依存吸引力上位 1 語をメイン・トピック，出現した他のすべての語をサブ・トピックとしてトピック遷移パターンを抽出した．

図 5，図 6 に生成した 2 種類の連鎖構造例を示す．たとえば図 6 では初めの語「電池」から終わりの語「問題」までを 3 つの場面を遷移して連結した．初めの語である「電池」が一番左上に示され，終わりの語「問題」が一番右下に示される．太いリンクはトピック遷移を表し「電池」からトピック遷移でつながる「展開」「アンテナ」「問題」という語が 3 つの場面のメイン・トピックを示している．3 場面のメイン・トピック「展開」「アンテナ」「問題」から細いリンクでつながる語は，各場面における遷移吸引力の上位 10 語を示している．

図 5 ではなるべく [M to M] でつなげて「問題」がサブ・トピックとして出現したときに終了させた．「電池」以外のトピックには特に留意せずに連結した．この連鎖構造からは「電池」を中心とする Tips を収集することができた．「電池」によるキーワード検索結果から適当な文書を絞り込んだものと変わらない結果となったが，この連鎖構造を後から参照することによって「電池」と「何」に関する文書であるかということを理

解するための支援ができるのではないかと考えた．

図 6 ではなるべく [S to M] でつなげて「問題」がメイン・トピックとして出現したときに終了させた．「電池」から「アンテナ展開」「プランジャ」「構造系」「問題」とトピックを遷移し，話題のつながりという点では一番もっともらしい連鎖構造を生成できた．この連鎖構造の 3 番目の場面には「電池」という語は含まれておらず，単なるキーワード検索では得難い結果が得られた．

本システムとのインタラクションを通じてグラフ構造を生成していくことにより，情報アクセスの支援を行うことができる．また，場面の結合方針を変化させることにより，場面の意味を多面的に捉えつつ文書群から横断的に潜在する複数の文脈を抽出することができる．

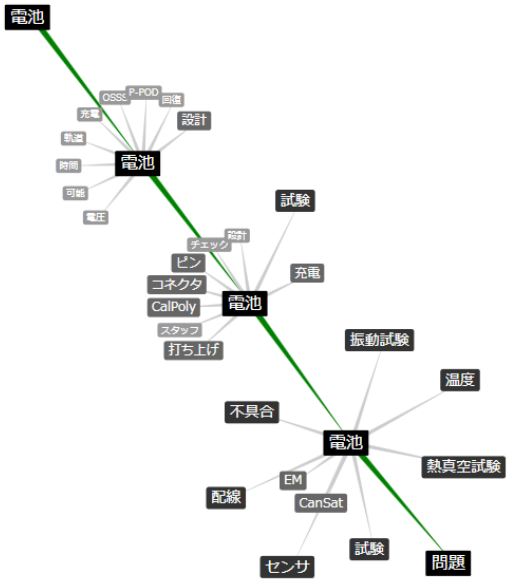

図 5: 連鎖構造例 1

5. おわりに

本稿では物語構造モデルに基づき話題の遷移を分析する手法を提案した．これは文書断片の構成によって文書断片内の語の関係や重要度が変化すると仮説を立てて，話題の遷移を定量的に評価する文脈依存吸引力を定義し実装した．

本研究の成果は以下の 3 点にまとめられる．

- 情報断片の意味を動的に扱う
- 情報アクセスを支援する
- 文書群に潜在する複数の文脈を抽出する

今後の課題としては以下の 2 点が挙げられる．まず，本研究で提案した文脈依存吸引力は鎖状構造の連鎖履歴のみを考慮した語の重要度の指標であるため，より複雑な場面構成を扱うためにはその影響を考慮した再定義を行う必要がある．次に，連鎖構造を生成するためには，妥当性と意外性のトレード・オフを考慮し，有用性というユーザの観点で候補場面を動的に定量評価する手法が必要である．

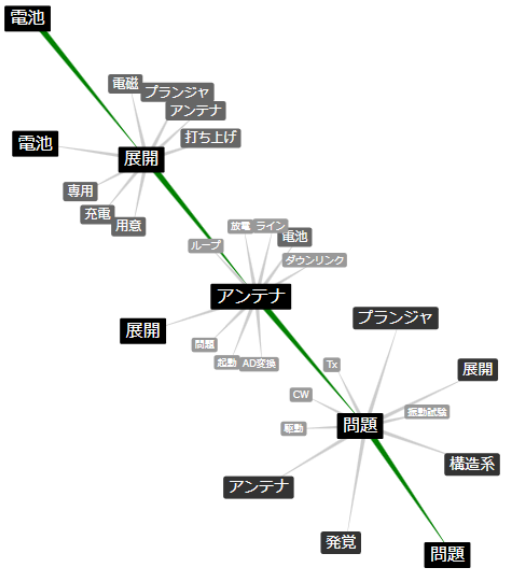

図 6: 連鎖構造例 2

参考文献

[Anderson 83] Anderson, J.: A spreading activation theory of memory, *Journal of Verbal Learning and Verbal Behavior*, Vol. 22, No. 3, pp. 261–95 (1983)

[Salton 75] Salton, G., Wong, A., and Yang, C. S.: A Vector Space Model for Automatic Indexing, *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620 (1975)

[佐藤 07] 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, 第 21 回人工知能学会全国大会論文集 (2007)

[赤石 06] 赤石 美奈：文書群に対する物語構造の動的分解・再構成フレームワーク, 人工知能学会論文誌, Vol. 21, No. 5, pp. 428–438 (2006)