論 文 Technical Papers

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

A Dynamic Decomposition/Recomposition Framework for Documents based on Narrative Structure Model

赤石 美奈
Mina AKAISHI
東京大学・先端科学技術研究センター
RCAST, University of Tokyo
akaishi@ai.rcast.u-tokyo.ac.jp

keywords: narrative, term dependency, term attractiveness, articulation, information retrieval

Summary

This paper proposes a framework to access information based on a narrative structure of documents. This framework consists of two processes. The one is to decompose existing documents into smaller units. The other process is combining unit components into a new story taking on a new meaning based on a context.

In this paper, a narrative structure for documents is modeled as follows. A story corresponding to a document is regarded as a sequence of scenes. A scene is a chunk of sentences. A sentence is mapped into a set of terms in the sentence. Decomposition process gives two mechanisms to decompose a story into scenes. Composition process shows four patterns to connect scenes. Both techniques to decompose/compose a story are based on the notions of *term dependency* and *term attractiveness*.

This paper also showes visualization tools to express the narrative structure for documents. *Word Colony* overviews content of a story as a directed graph representing the relation among *term dependency*. *Topic Sequence* is also directed graph to show the sequence of scenes along a story plot. The basis of these visualization techniques is the notions of *term dependency* and *term attractiveness*. They show the variety of understandings of the same documents.

1. はじめに

近年，様々な分野において，電子化された膨大な情報が蓄積されており，これを有効に利用するための具体的な手法が求められている．

蓄えられた情報を取り出すための代表的な技術として，様々な検索エンジンの研究開発が進められており，単語検索から全文検索へとその技術は進歩している．また，関連語や類似語等，いろいろな関連の知識を検索することも可能となってきている．検索される膨大な情報を的確に絞り込むことは困難な場合も多々生じるが，蓄積された情報の中から，条件に適するものを正確に高速に得る場合には，既存の検索エンジンは，強力な情報探索ツールであるといえる．

また，データマイニング，テキスト・マイニングなど，大量データから新たな法則性を発見するのに役立つ研究やツールの開発も，国内外において多々行われている．マイニングは膨大なデータから自動的に未知の法則を見つけ出すのに役立つツールであり，ある程度の発生頻度があり，高い信頼性のある規則が抽出される．

これらの技術は，蓄積された情報を静的なものとしてとらえ，蓄えられた情報の中から条件に合う情報や規則を取出すことを可能としている．これに対して，状況や文脈に応じて，動的に再構成された情報の必要性・重要性が広く認識されてきている．つまり，既に蓄積されている情報や規則を取り出すだけではなく，情報が必要とされている文脈に応じて，既存の情報を分解・再構成する技術の研究開発が必要とされている．

知識管理の分野においては，多視点から現象を分析し，動的に知識を創造する知識創造の必要性が指摘され，知識創造過程に関する理論が生まれている [野中 95, Fischer 01]．しかしながら，データに対する，多様な視点，柔軟なものの見方の必要性・重要性は，認知されているが，具体的な方法の研究は未だ充分とはいえない．

情報の意味のネットワークを多層的，立体的，多義的に構築するためには「分節」の技術が必要不可欠である．この分節のプロセスは，情報の分解と再構成の対により成り立つ．堀らは，これを「知識の液状化と結晶化」[Hori 96, Hori 05, 網谷 05] と呼び「知識の液状化」は「知識の結晶化」と対になり，知識創造を支えるプロセスであ

ると述べている．

人間は，多くの情報から，必要な箇所を抜き出し，繋ぎ合せ，状況に応じた文脈に沿って，新たな情報を生成することができる．本研究では，大量に蓄積された情報を機械的に処理して，大量の情報の中に隠された潜在的な物語を紡ぎ出すことを目標とする．このために，物語構造モデルを導入し，文書の意味を解釈せずに，文書から得られる表層的な特徴量を基に，物語構造を抽出し，文書を分解・再構成する，ナラティブ連想情報アクセス・フレームワークを提案する．

本論文は，以下のように構成される．2 章において，ナラティブ連想情報アクセスについて述べ，3 章において，本研究で基本概念として用いる「語の出現依存度」と「語の吸引力」について説明し，テキストから語彙連鎖グラフへの変換について述べる．この基本概念に基づき，4 章では，物語を分解する方法，5 章では，物語を再構成する方法，及び，ナラティブ連想情報アクセス・システムについて述べる．6 章において，まとめと今後の課題について述べる．

2. ナラティブ連想情報アクセス

物語に関する研究は，多々行われてきている．近年，知識を伝達するための物語の重要性が認識されてきている [Brown 05, Bringsjord 00, 松岡 96]．個別の知識だけを伝達するのではなく，文脈とともに伝達することで，より正確に情報が伝わることが報告されている．

既存の情報検索システムでは，指定された条件を満たす情報をユーザに提示する．しかしながら，情報は，蓄積された (記述された) 時に想定されていた文脈以外の文脈において有用な場合がある．情報を記録した時の文脈と，情報を利用する時の文脈は同じとは限らない．また，特定の情報を含みうる文脈を事前にすべて想定しておくことは不可能である．そこで，利用時の文脈に基づき，動的に情報を再構成して提示する技術が必要であると考えた．

本章では，ナラティブ連想情報アクセス・フレームワークとその基盤となる物語構造モデルに関して説明する．

2·1 文書に対する物語構造モデル

現代物語論において，G. ジュネット [Junet 85] によれば，物語は，以下の 3 つの側面を持つとされる．

- 物語内容 (語られた出来事の総体)
- 物語言説 (発話/記述された言説)
- 語り (語るという行為そのもの)

本研究は，物語言説から得られる特徴量を用いて，物語内容の構造を抽出し，その枠組みに基づき，物語の分解・再構成を行うものである．大量の文書を高速に処理し，新しい文脈に沿う情報生成支援を可能とするためには，テキストに対して，機械的に分節を行うことが必要である．このため，本研究では，言葉の意味や物語の内容を解釈せずに，物語の構造と文書の語彙連鎖構造のみに着目して分節を行う仕組みの研究・開発を目指している．

表 1 物語構造モデルと文書構成要素．

world model(世界構造)	set of stories
story (物語)	sequence of scenes (document)
scene (場面)	chunk of event
event (出来事)	set of terms (sentence)
character (登場人物)	term

さらに，松岡 [松岡 96] は，物語の構成要素を，

- ワールド・モデル (世界構造)，
- ストーリー (スクリプト，プロット)，
- シーン (場面)，
- キャラクター (登場人物)，
- ナレーター (語り手)

としている．

現代物語論では，語り手の視点に基づき，物語が展開されるため，語り手は物語の重要な要素と考えられている．これに対して，本研究で提案する物語構造モデル (表 1) は，文書の分解・再構成のためのモデルであり，物語の内容の解釈をしないため，語り手をモデルに含めていない．しかしながら，語りは，物語を語るための視点の役割を果たすものであり，本研究で提案するフレームワークにおいては，2·2 節で示す図 1 における composition rules(scene 結合規則) が，これに相当する．

これらの観点に基づき，文書の分解・再構成の基本とする物語構造モデルの構成要素と文書の構成要素との対応を表 1 に表す．

文章を構成する要素は，記号，文字，単語，文，段落などに分けることが可能である．ここでは，文章の最小構成要素を単語 (term) とし，これを登場人物 (character) に対応させる．また，登場人物の集合を扱うための単位として，登場人物が繰り広げる出来事 (event) の概念を導入する．これには，語の集合である文が相当する (データ構造としては，term の集合として扱う)．次に，まとまりのある一連の出来事 (event) により，場面 (scene) が構成され，場面の連結により物語 (story) が構成されると考える．この物語 (story) が，文書に相当する．さらに，物語の集合 (文書集合) により，対象とする世界構造 (world model) を規定することになる．

この時，scene の抽出方法により，異なる分節が可能となる．章，節，段落などを scene に対応させ，記述順序に沿い連結させることにより，文書の著者が想定した文脈に基づく story(元文書) が得られる．異なる分節を行えば，異なる文脈に基づく story が形成される．本研究においては，文書集合の中に埋もれている story を見つけるために，語彙連鎖に基づく連想を支援するための scene の抽出，結合を実現している．characer と event

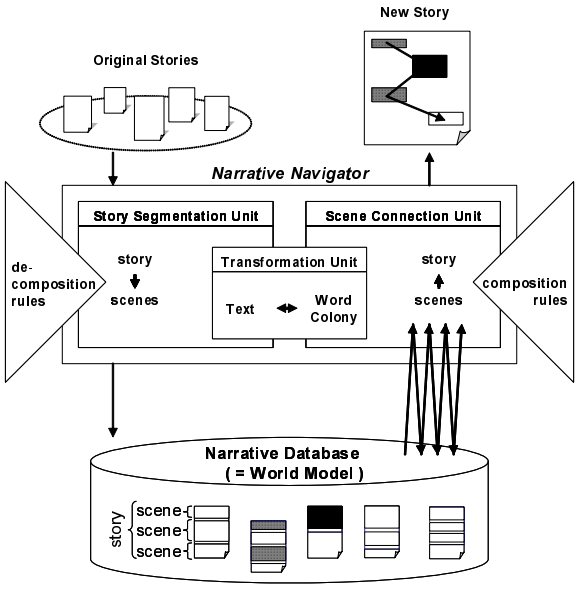

図 1 ナラティブ情報アクセス・フレームワーク .

の集合から導出される語彙連鎖の基本概念について 3 章で述べ , 語彙連鎖に基づく scene の抽出方法については 4 章で , 語彙連鎖に基づく scene 結合方法については 5 章においてそれぞれ述べる .

2·2 ナラティブ連想情報アクセス・フレームワーク

図 1 に , ナラティブ連想情報アクセス・フレームワークを示す . ナラティブ連想情報アクセスを可能とする Narrative Navigator (NaNa) は , 文書を分割する Story Segmentation Unit (SSU) と分割されたものを再構成する Scene Connection Unit (SCU) から構成される .

膨大な記録や資料を有効に利用するためには , 利用する際の文脈に沿った情報提示が必要である . 情報生成時と利用時の文脈は , 同じとは限らない . また , ひとつの文書には , 複数のトピックが文脈に沿って記述されている . このため , まず , 新しい文脈にあわせて情報を再構成するために , 元の情報を適当な粒度に分割するための文書分割技術が必要である (SSU) . 分割のための方法は , de-composition rules として与えられる . また , 分割された情報を新しい文脈に沿い , 結合させるための結合規則は , composition rules として与えられる . Narrative Navigator は , これらの二つの技術を組み合わせ , 文脈に沿った情報提示を行うことを可能とする .

本研究では , SSU が物語 (story) を場面 (scene) に分解し , SCU が場面 (scene) を連結し物語 (story) を形成する役割をもつ . また , SSU 及び SCU は , 線形に記述されたテキストから , 語彙連鎖関係を視覚化した有向グラフ Word Colony へ変換する Transformation Unit(TU) と連携し , グラフの分解・再結合を基にして , 語彙連鎖に基づく物語の分解・再構成機能を実現している .

3. 語の依存性に基づく語彙連鎖構造

本章では , 線形に記述された文章から , 語彙連鎖関係に基づくグラフ構造への変換に関して述べる .

本研究は , 大量に蓄積された文書集合を対象とし , それに対するナラティブ連想情報アクセスを可能とすることを目指している . 大量の情報を処理するために , 文書の内容を解釈せずに , テキストの表層から抽出される特徴量を基に , 語と語の連鎖関係を表すグラフ構造に変換し , グラフの構造に対する操作を通じて , 元の文書群の分解・再構成を実現している .

本章では , そのための基本概念 , 及び , 語彙連鎖グラフ (Word Colony) について説明する .

(なお , 簡単のため , 本論文で扱う例においては , 語として名詞のみを扱うこととする . また , テキストからの名詞の抽出には , 日本語形態素解析ツール茶筌 [Matsumoto 00] を用いている .)

3·1 語の出現依存度と吸引力

本節では , 文書に出現する「語の出現依存度」と「語の吸引力」について説明する .

文書に含まれるすべての語の集合を T とする . 文書中の異なる二語 , 語 $t \in T$ と $t' \in T$ に関して , 語 t から t' への出現依存度とは , 語 t が出現した同じ文中に語 t' が出現する確率 (条件付確率) と定義する . つまり , 文書において , 語 t の t' に対する出現依存度 $td(t,t')$ は , 以下の式で計算される .

$$td(t,t') = \frac{sentences(t \cap t')}{sentences(t)}, \qquad (1)$$

ここで , $sentences(t)$ は , 文書中における語 t を含む文の数であり , $sentences(t \cap t')$ は , t と t' を同時に含む文の数である .

次に , 語 t が文書中の他の語を引き付ける力を吸引力と呼び , 他の語から語 t に対する出現依存度の総和として , 以下のように定義する .

$$attr(t) = \sum_{t'' \in T} td(t'',t), (\text{ただし }, t \neq t'') \qquad (2)$$

3·2 語彙連鎖に基づく主題俯瞰:Word Colony

Word Colony [Akaishi 04a, 赤石 04b, 赤石 05] は , 文書中の語の出現依存関係の方向性に着目し , 語群クラスターを形成し , 文書の内容を語と語の関係として視覚化するツールである . 大量の情報が氾濫している状況において , 興味のあるテーマに関する文章すべてに目を通すことは不可能である . このため , 文書中の重要文を抽出して , 自動的に要約を生成する技術が必要とされ , 研究・開発されている [Inderjeet 03] . これに対して , 文書の語の共起関係を視覚化した Word Colony は , 視覚的要約と捉える事ができる .

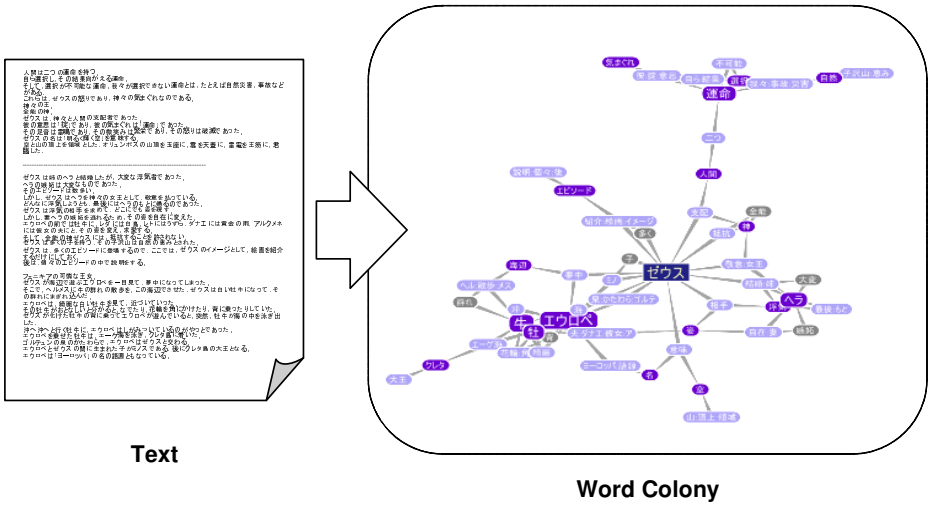

図 2 語彙連鎖グラフ Word Colony．

文章の内容を語の共起性に基づきグラフとして視覚化するツールとしては，KeyGraph[Ohsawa 98, 大澤 99] が提案されている．KeyGraph では，キーワード抽出のための語の重み付け手法として，頻出語と強い共起関係にある語に着目し，従来の方法では見落とされていた重要な語を抽出することを可能としており，チャンス発見の基本ツールとしても利用されている．KeyGraph の基本となる語の共起関係には，方向性はなく，出現頻度の低い語同士の共起関係や，頻出語と弱い共起関係にある語は，グラフには表れない．

しかしながら，出現頻度の低い語や，弱い共起関係にある語も，文脈によっては，重要な語となる可能性がある．Word Colony では，語の吸引力は，他の語との出現依存関係に基づき定義されているため，出現頻度の低い語は，その語が依存している他の語の吸引力を強くするために貢献しているという特徴がある．また，方向性を考慮した語と語の出現依存関係を用いることにより，共起関係の強弱も表現されている．

各語間の出現依存度を指標として用いることにより，文書に出現する二つの語の間には，(i) 双方向に強い依存度を持つ場合，(ii) 一方向にのみ強い出現依存度を持つ場合，(iii) どちらの方向に対しても低い依存度を持つ場合の 3 つの場合が考えられる．Word Colony は，これらの関係を以下の様に視覚化する．

(1) 双方向に強い出現依存度を持つ語
二つの語 t と t' が，双方向に強い出現依存度を持つ場合，それらの語は，その文書において，非常に密接に関係していると考えられる．そこで，互いの出現依存度が閾値 σ を超える場合 $(td(t,t') > \sigma$ かつ，$td(t',t) > \sigma)$，それらの語は，同一グループに属するものとみなし，ひとつのノードにまとめる．各ノードの大きさは，ノードに含まれる語の吸引力の和に比例する[*1]．(ただし，$0 \leq \sigma \leq 1$)

(2) 一方向に強い出現依存度を持つ語
二つの語 t と t' において，語 t から t' への出現依存度が閾値 δ より大きく，かつ，語 t' から t への出現依存度が閾値 μ より小さい場合 $(td(t,t') > \delta$ かつ，$td(t',t) < \mu)$，語 t のノードから，語 t' のノードへリンクを張る．つまり，依存する語ノードから，依存される語ノードへリンクが張られる．(ただし，$0 \leq \mu \leq \delta \leq 1$)

(3) 出現依存関係のない語
上記の条件を満たさない語 t と t' の間には，出現依存関係はないものとみなす．

図 2 に，ギリシャ神話の「エウロペ」に関する話を題材とした文書 (図 2 左上) を，Word Colony で視覚化した結果を示す[*2]．この文書に登場する語 (例:ゼウス，エウロペ，ヘラ，牡牛など) の依存関係に基づき，文書の内容が有効グラフとして視覚化されている．

4. 語彙連鎖に基づく story 分割

ひとつの文書には，複数の主題が含まれている．本章では，語の出現依存度と語の吸引力に基づき，各 scene のメイン・トピックとなる語の吸引力を最大にするように文書を分割する方法について述べる．この時，元の文書に書かれた文章の連続性を保持したまま文書を分割する「主題遷移解析に基づく系列的分割法」と，文章の連続性を保持せずに分割する「主題階層解析に基づく場面的分割法」について述べる．

4·1 主題遷移解析に基づく系列的分割法

本節では，文書に記述された文章の順序に沿い，語の吸引力の強さの変化に着目し，scene を抽出する手法につ

*1 本論文では，特に断らない限り，語の吸引力が最大である語をメイン・トピックと呼び，Word Colony においては，矩形ノードで示す．また，語の吸引力が最大である語が複数ある場合は，それら複数の語をメイン・トピックとする．

*2 この例では，$\sigma = \mu = \delta = 0.9$

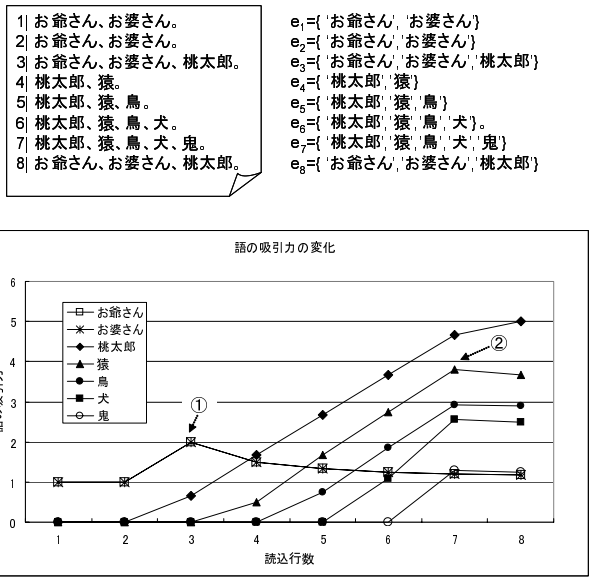

図 **3** 語の吸引力の変化 .

いて述べる . これは , 着目した語の吸引力が極大になる箇所を , 文章内のトピックの変化箇所として検出し , 文書を分割する手法である .

まず , 文書 D に表れる語の集合を $T = \{t_j|1...m\}$(ただし , m は , D に含まれる異なり単語数) とする . このとき , event は , 文書 D の各文に含まれる語の集合で表される . (図 3 左上の例文に対する event , e_1, e_2,... を図 3 右上に示す .)

ここで , 文書 D における event の集合を $E = \{e_i|i = 1...n\}$(ただし , i は行番号 , n は D 内の文章数に相当) , 集合 E の部分集合 E_i を $E_i = \{e_k|k = 1...i\} \subset E$(ただし $n \geq i \geq 1$) としたときに*3 , 各 E_i における語 t_j の吸引力を $attr_i(t_j)$ とする . i を変化させたときの $attr_i(t_j)$ の変化は , 文書内の文章を順番に読み込んだときの語 t_j の吸引力の変化を表す . 例えば , 図 3 の例文を用いると , 語"桃太郎"の吸引力は , 以下のように計算される (ただし , 依存度が 0 である項は省略) .

$$\begin{aligned} attr_3(\text{"桃太郎"}) &= td_3(\text{"お爺さん","桃太郎"}) \\ &+ td_3(\text{"お婆さん","桃太郎"}) \\ &= 1/3 + 1/3 = 2/3 \\ attr_5(\text{"桃太郎"}) &= td_5(\text{"お爺さん","桃太郎"}) \\ &+ td_5(\text{"お婆さん","桃太郎"}) \\ &+ td_5(\text{"猿","桃太郎"}) \\ &+ td_5(\text{"鳥","桃太郎"}) \\ &= 1/3 + 1/3 + 2/2 + 1/1 = 8/3 \end{aligned}$$

図 3 下に , 例文で用いられた各語の吸引力の変化をグラフで示す . 語の吸引力が変化しないか , 減少する部分を停滞部 , 増加する部分を上昇部とすると , 読込行数 i を変化させることにより , 語の吸引力は , 停滞部と上昇部を交互に示す . ここで , 上昇部は , 着目した語がトピックとして語られている部分と解釈できる . そこで , 上昇部から停滞部へ変わる箇所において , 着目した語のトピックとしての成長が局所的に止まったと解釈し , 文書の分割箇所として検出する . 図 3 の例においては , "お爺さん"に着目した場合 , 読込行数が 1 行から 2 行までが停滞部 , 2 行から 3 行までが上昇部 , 3 行から 8 行までが停滞部となる . この 3 行目を"お爺さん"で表されるトピックの終了とみなす . また , "猿"に着目した場合は , 7 行目がトピック"猿"の終了箇所となる . 系列的分割箇所検出のアルゴリズムの詳細は , 付録 A に示す .

図 4 は , 図 2 で用いたギリシャ神話の文書に対して , 主題遷移解析を行い , 文書 (story) を scene に分割した例である . 図 4 左には , 各 scene 毎に Word Colony にして , 上から下につないだ Topic Sequence が示されている . それぞれの scene におけるメイン・トピックが「運命」「ヘラ」「姿」「ゼウス」「エウロペ」「エウロペ : ゼウス」「ヨーロッパ:エウロペ」と変化しており , それぞれの主題毎に scene 分割がなされている .

4·2 主題階層解析に基づく場面的分割法

本節では , 文書に記述された文の順序に捉われず , 語の吸引力の強さに着目し , scene を抽出する手法について述べる .

Word Colony は , 線形に記述されたテキストから , 語と語の依存関係だけに着目し , 視覚化したものである . 生成された Word Colony から , 吸引力の強い語 (メイン・トピック) の影響を排除することで , 他の語同士の隠れていた依存関係が顕在化される . これを利用して , 文書を分割する手法を主題展開に基づく場面的分割法と呼ぶ .

文書から生成された Word Colony に含まれる連結成分をトピック・クラスターと呼ぶ . このトピック・クラスターに含まれている語は , 出現依存関係により連繋している語のグループである . この時 , 吸引力の強い語を削除して , Word Colony を再生成するにつれ , グラフが分解されていく . 細分化された Word Colony のトピック・クラスターに含まれている語を含む文のみを , 元の文書から抜き出すことにより , ある主題に関する event を集めた scene を構成することができる .

まず , 元の文書に対する Word Colony をグラフ WC_0 とした時に , WC_0 に含まれるトピック・クラスターからなる集合を TP_0 とする . ただし , 各トピック・クラスターに含まれるノード数は , パラメータ min_n (≥ 1) より大きいものとする . 次に , WC_0 における吸引力が最大の語を削除し , 再生成した Word Colony を WC_1 とし , これに対するトピック・クラスター集合を TP_1 とする . これを i 回繰り返した時のトピック・クラスター集

*3 $E_1 = \{e_1\}, E_2 = \{e_1, e_2\}, E_3 = \{e_1, e_2, e_3\}$,...

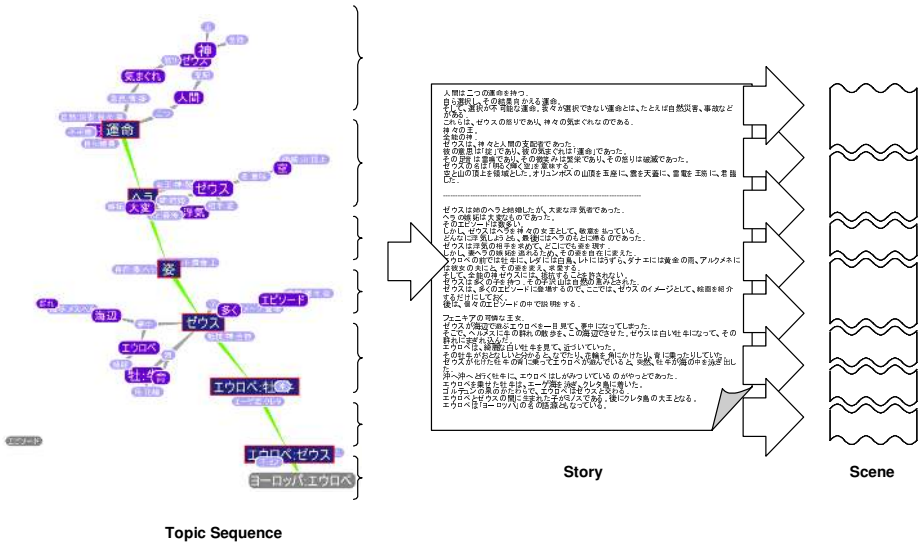

図 4 系列的分割例．

合を TP_i とし，この要素の個数を $n(TP_i)$ とする．i を変化させた時に $n(TP_i)$ が最大となる時に，トピック・クラスターが最細分されたとみなし，それぞれのトピック・クラスターに含まれる語を含む文 (event) を抜き出し，scene を抽出する．

図 5 は，図 2 で用いたギリシャ神話の文書に対して，次々とメイン・トピックを消去し，再計算した Word Colony WC_1，WC_2，WC_3，WC_4，WC_5 を示している．各段階でのメイン・トピックを図中に示す．元の文書に対する Word Colony WC_0 のメイン・トピックは「ゼウス」である (図 2 参照)．吸引力の一番強い語「ゼウス」を消去し，再計算をして Word Colony を生成することにより，「ゼウス」の影響を排除した Word Colony WC_1 を生成できる．WC_1 において，メイン・トピックとして表れたのは「牛」である．これを排除し，同様の操作を繰り返すことにより，図 5 の Word Colony WC_5 が得られる．WC_5 の各トピック・クラスターに含まれる語を含む文章を抜き出した結果，「ゼウスの神としての性質」について記述された部分や，「ゼウスの浮気」に関する記述，「ゼウスやエウロペの名前の由来」に関する記述部分などの scene が，テキストの記述順序に捉われず抜き出される．

5. 語彙連鎖に基づくナラティブ連想

4 章では，語の出現依存関係と吸引力に着目し，文書を分解する方法について述べてきた．本章では，語の出現依存関係に着目して文脈を生成しながら，scene をつなげて story を生成する，scene 結合規則について説明し，2·2 節にて述べたナラティブ連想情報アクセス・フレームワークに基づき，ユーザの情報アクセスを支援する Narrative Navigator (NaNa) に関して述べる．

連想的文書検索システムとしては，DualNavi [Takano 00a, Takano 00b, Takano 03] が提案されている．これは，類似文書検索と特徴語グラフを有機的に連携させたシステムであり，連想的情報アクセスの有用性を示している．DualNavi では，対象文書に含まれる語の共起に基づき，検索される類似文書が結果として出力される．

これに対して，本研究で提案するナラティブ連想情報アクセスは，対象文書を，物語構造に基づき分節し，ユーザが選択した文脈に沿って生成される新しい story の候補を結果として出力し，既存文書集合を横断的に再構成して得られる新しい知識獲得を支援するものである．

5·1 場面 (scene) 結合規則

本節では，主題遷移に着目し，scene をつなげて文脈を生成しながら，story を形成するトピック遷移パターンについて述べる．

図 6 に，メイン・トピックの遷移パターンの模式図を表す．scene の内容は，Word Colony で表しており，語をノードで表し，その吸引力をノードの大きさで表現している．パターン [M to M] は，ある scene のメイン・トピックが，次の scene でもメイン・トピックになっている場合，パターン [S to M] は，ある scene のサブ・トピックが，次の scene のメイン・トピックになっている場合，パターン [New M] は，ある scene には出現しない語が，次の scene ではメイン・トピックになっている場合である．

つまり，ある scene を $scene_i$ とし，続く scene を $scene_{i+1}$ とし，それぞれの scene におけるメイン・トピック (吸引力の強い語) の集合を M_i，M_{i+1}，サブ・トピック (吸引力の弱い語) の集合を S_i，S_{i+1} とした時に，$scene_{i+1}$ における任意のメイン・トピック $m_{i+1} \in M_{i+1}$ が，前の $scene_i$ においてメイン・トピックであったか，サブ・トピックであったか，出現していなかったかの 3 つのパターンである．

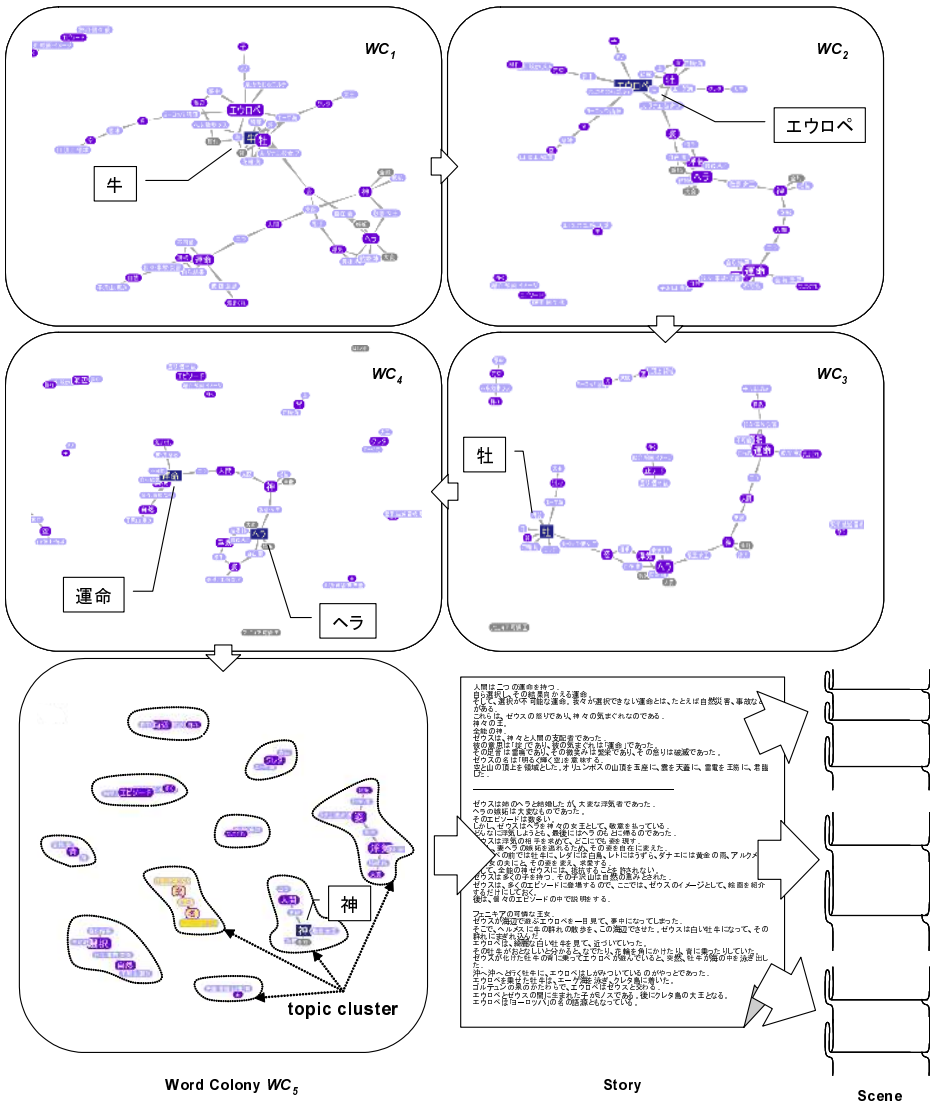

図 5 場面的分割例 .

表 2 トピック遷移パターン .

メイン・トピック遷移パターン	
[M to M]	$\exists m_{i+1} \in M_{i+1},\ m_{i+1} \in M_i$
[S to M]	$\exists m_{i+1} \in M_{i+1},\ m_{i+1} \in S_i$
[New M]	$\exists m_{i+1} \in M_{i+1},\ m_{i+1} \notin M_i \cup S_i$

サブ・トピック遷移パターン	
[M to S]	$\exists s_{i+1} \in S_{i+1},\ s_{i+1} \in M_i$
[S to S]	$\exists s_{i+1} \in S_{i+1},\ s_{i+1} \in S_i$
[New S]	$\exists s_{i+1} \in S_{i+1},\ s_{i+1} \notin M_i \cup S_i$

サブ・トピックの遷移も同様の 3 パターンに分けられる . これらのトピック遷移パターンとその条件を表 2 にまとめて示す .

5·2 語彙連鎖に基づく文脈生成

本節では , 前節で述べたトピック遷移パターンの出現割合に関して述べ , ナラティブ連想に必要な語彙連鎖パターンについて考察する .

図 7 は , 短編小説 2 編 (芥川龍之介の「羅生門」と「蜘蛛の糸」) と合成テキスト (「羅生門」と「蜘蛛の糸」の第 1 段落から第 13 段落までを交互に並べたテキスト) を例として , 各段落の連結部でのメイン・トピックの遷移パターンの出現割合を調べたものである*4 .

オリジナルの物語である「羅生門」と「蜘蛛の糸」において , メイン・トピックが遷移パターン [M to M] で遷移している連結部の割合は , 12% , 33% であり , [S to M] で遷移している割合は , 40% , 42% であった . これらは , 各 scene のメイン・トピックが , あらかじめ前の段落で出現しており , それを継続 , あるいは伏線として徐々に主題を遷移させていることを示す . また「羅生門」「蜘蛛の糸」それぞれにおいて , メイン・トピックが , 遷移パターン [New M] で遷移する割合は , 約 48% と約

*4 吸引力が最大の語 (複数可) をメイン・トピックとし , それ以外の語は , サブ・トピックとした . メイン・トピックが複数ある場合は , 複数の遷移パターンが生じる場合があり , 各パターンの出現割合の合計は 100% にならないことに注意 .

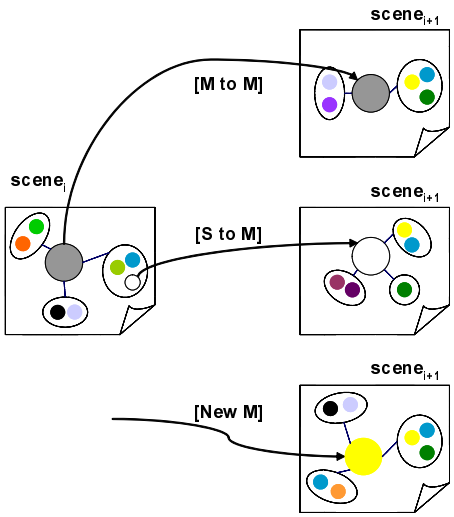

図 6　ナラティブ連想パターン (メイン・トピック遷移パターン)．

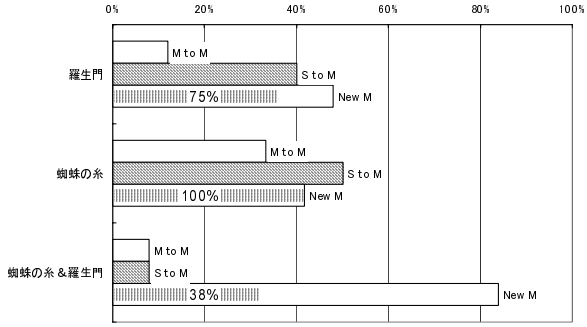

図 7　語彙連鎖結合パターンの出現割合．

41% であった．これらは，前の scene では，出現していない語が，メイン・トピックとして表れる割合を示している．

これに対して，合成テキストでは，遷移パターン [M to M]，及び [S to M] の割合が，それぞれ 8% であり，[New M] の割合が，84% であった．

これらから，内容を無視して連結された合成テキストに比べて，人間によって記述された物語においては，遷移パターン [M to M]，及び [S to M] で連結されている割合が大きいという特徴が明らかである．

このことより，story を形成する scene の連結においては，[M to M]，あるいは，[S to M] の条件を満たす候補を優先的に提示することで，ユーザが，もっともらしい自然な文脈を生成することを支援できると考えられる．

また，[New M] パターンで遷移している場合に，サブ・トピック遷移パターン [M to S] か [S to S] が表れている割合は，「羅生門」では 75%，「蜘蛛の糸」では 100%，「合成テキスト」では 38% であった．つまり，人間によって記述された物語においては，新しい語がメイン・トピックとして出現する場合にも，高い確率でサブ・トピックが前の段落の語から遷移しており，scene との関連を表していることがわかる．

ここで，[New M] の遷移に関しては，サブ・トピックの遷移 [M to S] か [S to S] を候補とすることで，新しいメイン・トピックの候補を示すことが可能である．しかしながら，通常，メイン・トピックは，1〜数個であり，サブ・トピックは数十個以上であることから，[M to S] か [S to S] の条件で絞り込んだとしても，候補の数が膨大になるのは容易に推測される．これらの候補に関しては，語や文章の意味を考慮し，概念辞書などを利用した重み付けなどを用いて，さらに候補を絞り込む仕組みが必要がある．しかしながら，この点に関しては，本論文の範疇を外れるため，ここでは論じない．

以上より，本システムは，ユーザの連想を促すための連想パターンとして，[M to M]，あるいは，[S to M] の連結を網羅的に探索し，可能な story をユーザに提示し，ナラティブ連想情報アクセスを支援することとする．文書から分割された scene をこのパターンで連結することにより，表 1 に示した物語構造における story を生成することが可能となる．ただし，この特徴は，連結された scene が物語であることの必要条件とはなり得るが，十分条件ではないため，ユーザ自身が，システムが提示した連結候補から妥当な scene を選びながら情報にアクセスしていくことにより，そのアクセス過程からナラティブ連想パスを形成し，story を生成していくことになる．

5·3　Narrative Navigator

本節では，ギリシャ神話の物語 (60 話) を表 1 における世界モデルとして設定し，ギリシャ神話における「人」と「神」の関係を知りたいと仮定して，ナラティブ連想情報アクセスについて考察する．

本節では，図 1 における元文書集合として，ギリシャ神話の物語 (60 話) を設定し，NaNa を用いたナラティブ連想情報アクセスについて考察する．story の分解機構 (SSU) における分解規則 (de-composition rules) としては，4 章において述べた系列的分割規則と場面的分割規則を用いる．それぞれの方法により分解された scene の情報とともに，文書集合は，Narrative Database に格納される．

scene の再結合機構 (SCU) においては，5·1 節において述べたトピック遷移パターンを結合規則 (composition rules) として適用し，可能な scene 連結をユーザに示唆する．

WorldModel ノードは，ナラティブ情報アクセスの開始点となる．これは，表 1 に示した World Model に相当し，アクセスする文書集合を規定する．ここから，topic ノードを介して，ナラティブ DB 内に含まれるすべての scene ノードへリンクが張られる．ひとつの topic ノー

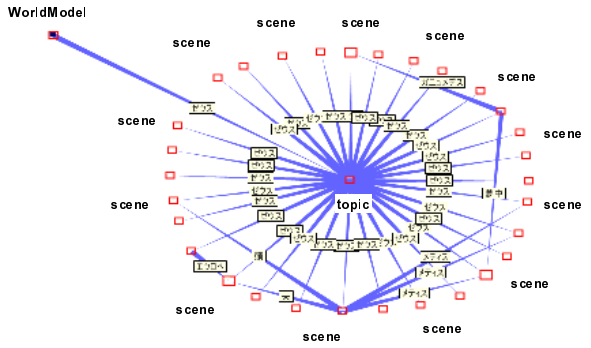

図 8 「ゼウス」を主題とする (系列的分割)scene からのナラティブ連想パス．

ドに繋がる scene ノードは，同じ語をメイン・トピックとする scene であり，この scene 間の連携は，5･1 節で述べたメイン・トピック遷移パターン [M to M] 連結に相当する．さらに，各 scene ノードからは，[S to M] パターンに基づき，他の scene ノードへリンクが張られており，ユーザが取捨選択をしながら情報アクセスした経過がナラティブ連想パスとして残され，リンクで連結された一連の scene により，story が生成される．

例えば，「ゼウス」に関して知りたいとする．図 8 では，WorldModel ノードにおいて，「ゼウス」の topic ノードが選択され，それに繋がるすべての scene ノードが示されている．これらは，同じ語「ゼウス」をメイン・トピックとする scene である．これらを連携することが，メイン・トピック遷移パターン [M to M] 連結に相当する．また，これらの scene 間の [S to M] 連結がいくつか示されている．この例では，複数の文書にまたがる scene を連結させて読み取ることにより，「メティスに夢中となったゼウス (アテナ物語の scene) が，ガニュメデスにも夢中になった (ガニュメデス物語の scene)」ことや，「メティスにクロノス攻略法を教えてもらったゼウス (ティタン戦争物語の scene) が，メティスを飲み込み (アテナ物語の scene)，ゼウスの頭から生まれたのが，メティスとゼウスの娘，知恵の女神アテナである (アテナ物語の scene)」ことなど，「ゼウス」の側面を表す story を生成できる．

次に，「人間」と「神」の関係についての情報を知りたいとする．まず，WorldModel ノードから，「人間」と「神」の Topic ノードを通じて，それぞれをメイン・トピックとする scene ノードへのリンクを辿る．さらに scene ノードから，[S to M] パターンを用いて scene 結合した過程が図 9 に示されている．情報アクセスは，図 9 の左から右方向に進められた．その結果，「神に罰を与えられた人間の女」や「乙女達を馬鹿にして罰を与えられた男」，あるいは，「神と人間が結婚するための条件から秘密にしている結婚」などの話を生成することができる．これらも，複数の文書から抽出された scene を横断して生成された story である．

また，図 10 は，図 9 と同じように，「人間」と「神」に

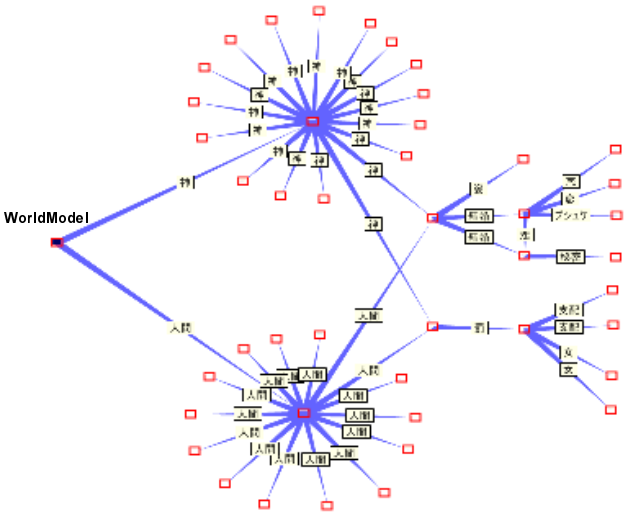

図 9 「人間」と「神」を主題とする (系列的分割)scene からのナラティブ連想パス．

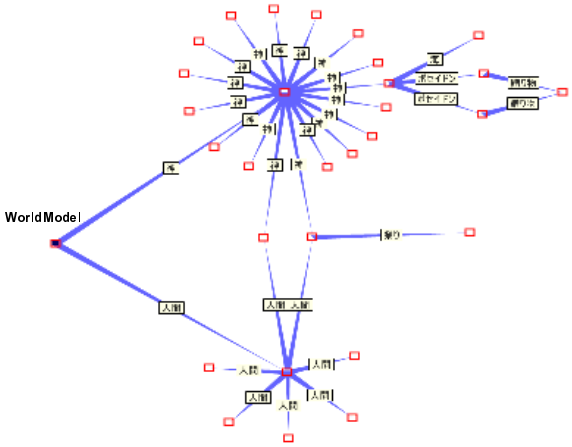

図 10 「人間」と「神」を主題として含む (場面的分割)scene からのナラティブ連想パス．

ついての関係を探った経過であるが，図 9 が，分割規則 (de-composition rule) として系列分割法を用いているのに対して，図 10 では，場面的分割法を用いているため，異なる scene が抽出され，抽出された scene に基づく語彙連鎖により story が生成されている．このため，図 9 の場合とは異なるナラティブ連想パスを形成している．

本論文においては，ナラティブ連想情報アクセスの基本概念とフレームワークについて説明した．対象の文書集合を規定する World Model は，アクセスする情報の範囲を規定する概念であり，対象文書集合に含まれる文書のドメインを限定することにより，生成される story の妥当性をある程度まで絞ることができる．ここでは，ギリシャ神話を例題として示したが，著者等は，本フレームワークを設計会議の議事録などに適用する実験も行っている．[赤石 06] においては，プロジェクトにおける機器の故障とその原因を，議事録の文書集合を横断する story

として見つけた事例について報告している．

6. おわりに

本論文では，物語構造に基づき，新しい文脈を生成しながら，必要な scene を連結し story を生成するナラティブ連想情報アクセスのフレームワークに関して述べた．

まず，文書における物語構造のモデルを定義し，文脈に基づき，情報を再構成するために必要な文書分割方法に関して，主題遷移解析に基づく系列的分割法と，主題階層解析に基づく場面的分割法を提案した．これにより，既存の文書を story とみなして，その中から意味のあるまとまりとしての scene を分割・抽出することを可能とした．また，scene を，語彙連鎖に基づくトピック遷移パターンに沿って結合することにより，複数の文書を横断する新しい文脈に基づく story 生成が可能なことを示した．

本研究では，物語の構造に着目し，語の意味や物語の内容は考慮していない．この特徴を生かして，テキスト以外の対象へと応用を広げるとともに，物語論研究における成果との連携について検討することが，今後の課題として残されている．

謝　辞

本稿執筆にあたり，本研究をはじめるきっかけと多様な示唆を頂いた北海道大学の田中譲先生，率直な批評と激励を頂いた国立情報学研究所の佐藤健先生，本研究の可能性を広げて頂いた東京大学の堀浩一先生，システム開発に多大な技術支援をして頂いた NorthGrid 社の菊池敏幸氏に心より感謝致します．

◇ 参 考 文 献 ◇

[Akaishi 04a] Akaishi, M., Satoh, K., and Tanaka, Y.: An Associative Information Retrieval based on the Dependency of Term Co-occurrence, *Lecture Notes in Artificial Intelligence*, Vol. 3245, pp. 195–206 (2004)

[赤石 04b] 赤石美奈, 田中譲, 佐藤健：共起依存度を用いた語彙連鎖に基づく連想的情報断片探索手法, 人工知能学会研究会資料、人工知能基礎論研究会, pp. 19–23 (2004)

[赤石 05] 赤石美奈, 堀浩一, 佐藤健：文書における語の共起依存性に基づく主題の視覚化, 人工知能学会研究会資料、人工知能基本問題研究会, pp. 89–94 (2005)

[赤石 06] 赤石美奈, 佐藤健, 堀浩一：語の吸引力に基づく主題遷移解析と視覚化, 電子情報通信学会技術研究報告、言語理解とコミュニケーション NLC2005-121 (2006)

[網谷 05] 網谷 重紀, 堀 浩一：知識創造過程を支援するための方法とシステムの研究, 情報処理学会論文誌, Vol. 46, No. 1, pp. 89–102 (2005)

[Bringsjord 00] Bringsjord, S. and Ferrucci, D. A.: *Artificial Intelligence and Literary Creativity*, Lawrence Erlbaum Associates (2000)

[Brown 05] Brown, J. S., Denning, S., Groh, D., and Prusak, L.: *Storytelling in Organizations*, Elsevier Butterworth-Heinemann (2005)

[Fischer 01] Fischer, G. and Ostwald, J.: Knowledge Management: Problems, Promised, Realities, and Challenges, *Intelligent Systems*, Vol. 16, No. 1, pp. 60–72 (2001)

[Hori 96] Hori, K.: A Model to Explain and Predict the Effect of Human-Computer Interaction in the Articulation Process for Concept Formation, *Information Modeling and Knowledge Bases*, Vol. 7, pp. 36–43 (1996)

[Hori 05] Hori, K.: Do knowledge assets really exist in the world and can we access such knowledge? - Knowledge evolves through a cycle of knowledge liquidization and crystallization -, *Lecture Notes in Artificial Intelligence*, Vol. 3359, pp. 1–13 (2005)

[Inderjeet 03] Inderjeet Mani(著), 奥村 学, 植田 禎子, 難波 英嗣 (翻訳)：自動要約, 共立出版 (2003)

[Junet 85] Junet G.(著), 花輪光, 和泉涼一 (訳)：物語のディスクール□方法論の試み, 水声社 (1985)

[Matsumoto 00] Matsumoto, Y., Kitauchi, A., Yamashita, T., Hirano, Y., Matsuda, H., Takaoka, K., and Asahara, M.: *Manual of Japanese Morphological Analysis System ChaSen version 2.2.1* (2000)

[松岡 96] 松岡正剛：知の編集工学, 朝日新聞社 (1996)

[野中 95] 野中 郁次郎, 竹内 弘高, 梅本 勝博：知識創造企業, 東洋経済新報社 (1995)

[Ohsawa 98] Ohsawa, Y., Benson, N. E., and Yachida, M.: KeyGraph: Automatic Indexing by Co-occurrence Graph based on Building Construction Metaphor, in *IEEE ADL'98*, pp. 148–156 (1998)

[大澤 99] 大澤幸生, Benson, N. E., 谷内田正彦：KeyGraph: 単語共起グラフの分割・統合によるキーワード抽出, 電子通信学会誌論文誌, No. 2, pp. 391–400 (1999)

[Takano 00a] Takano, A., Niwa, Y., Nishioka, S., Iwayama, M., Hisamatsu, T., Imaichi, O., and Sakurai, H.: *Information Access Based on Associative Calculation*, pp. 187–201, Springer-Verlag, sofsem 2000, lncs vol. 1963 edition (2000)

[Takano 00b] Takano, A., Niwa, Y., Nishioka, S., Iwayama, M., Hisamitsu, T., Imaichi, O., and Sakurai, H.: Associative Information Access Using DualNAVI, in *Kyoto International Conference on Digital Libraries (ICDL'00)*, pp. 285–289 (2000)

[Takano 03] Takano, A.: Association Computation for Information Access, in *Proc. of Discovery Science 2003, LNCS 2843*, pp. 33–44 (2003)

〔担当委員：阿部 明典〕

2005 年 10 月 3 日　受理

著 者 紹 介

赤石　美奈(正会員)

1995 年北海道大学大学院工学研究科電気工学博士課程修了．博士 (工学)．同年静岡大学工学部知能情報工学科助手．1997 年北海道大学工学研究科電子情報工学専攻助手．2004 年東京大学先端科学技術研究センター助教授．現在に至る．知識メディア，メディア・ベース，情報視覚化等の研究に従事．

◇ 付　　録 ◇

A. 系列的分割箇所検出アルゴリズム

系列的分割箇所検出のアルゴリズムの詳細を擬似コードとして以下に示す．

```
int [] segments(float attr[][], int m, int n,
                                int order_t, int min_s)
{
    int i, j, p=0
    int ttp[n];   /* ttp:分割箇所格納用配列   */
    ttp[0] = n;

    for (i=n; i>0; i--) {
        if ((ttp[p]-i)≥ min_s) {
            for (j=1; j≤m; j++) {
                if (attr[i][j] > attr[i−1][j] &&
                    attr[i][j] ≥ attr[i+1][j] &&
                    order(i, j) ≤ order_t ) {
                    ttp[++p] = i }
            }
        }
    }
    return ttp;
}
```

各語の吸引力は，既に計算され，二次元配列 attr[n+1][m+1] として引き渡されるものとする*5．引数 m，n は，それぞれ，文書 D に含まれる異なり単語数 m と行数 n である．

また，分割を制御するために二つの引数を導入する．一つの scene に含まれる最小 event 数を定義するための引数 $min_s(min_s > 0)$ と，分割の際に着目する語を特定するための引数 $order_t$ である．ここでは，$order_t$ は，着目する語の吸引力の最低順位とする (ただし，$order_t > 0$ であり，語の吸引力の順位とは，吸引力が最大のものを 1 位とする．)．

トピック変化場所集合は，配列 ttp に格納され，1 番目の要素として n が代入されている．分割箇所が検出されると，ttp に，分割行番号が追加されることになる．

読込行数 i を n から 1 まで変化させた時に，近傍分割箇所から，min_s 行以上離れた場合に，
(i)$attr_i(t_j)$ が $attr_{i-1}(t_j)$ より大きく，かつ
(ii)$attr_i(t_j)$ が $attr_{i+1}(t_j)$ 以上であり，かつ
(iii)$attr_i(t)$ の順位が，$order_t$ で指定された順位以下である
場合に，i の値をトピック変化場所として，配列 ttp に追加する．関数 $order(i,j)$ は，E_i における各語の吸引力を大きい方から降順に並べた場合に，語 t_j の吸引力 $attr_i(t_j)$ の順位を返す関数である．結果として，各 scene の終わりを示す event の位置 (行番号) が．ttp の要素として格納される．

つまり，読込行数 i を変化させた時に，語の吸引力が上位 $order_t$ 位までの吸引力を持つ語に着目し，吸引力の増加が止まり，減少に転じる箇所を話題の転換時点とみなし，文書を分割する手法である．

本文の図 3 の例題において，$min_s = 1$，$oreder_t = 1$ とした場合には，$ttp = \{8, 3\}$ となり，$min_s = 1$，$oreder_t = 2$ とした場合には，$ttp = \{8, 7, 3\}$ となる．

*5　説明の簡略化のため，配列の大きさを問題にせず，配列の添字と行番号，単語番号をそろえている．つまり，語 t_j の E_i における吸引力 $attr_i(t_j)$ は，attr[i][j] に格納されている．