様式 C-19

科学研究費補助金研究成果報告書

平成 22 年 6 月 2 日現在

研究種目：基盤研究(B)
研究期間：2006 ～ 2009
課題番号：18300046
研究課題名（和文）物語構造に基づく文書群の動的分解・再構成フレームワークに関する研究
研究課題名（英文） A Dynamic Framework for Re-construction of Documents based on Narrative Structure Model

研究代表者
赤石 美奈（AKAISHI MINA）
東京大学・先端科学技術研究センター・准教授
研究者番号：60273166

研究成果の概要（和文）：

本研究では、情報を伝達するための「物語」の重要性に着目し、文書(テキスト) における物語構造モデルを定義し、物語を構成する要素の関係に基づく文書の分節化の基本的手法を提案した。分節化とは、区切ることによって関係を生じさせることであり、分節化の具体的手法の研究開発は、既存の情報の粒度を変化させることにより、異なる関連を生み出し、文脈に応じて情報を再構成するために必要不可欠な技術である。本研究では、対象文書の中に出現する「語の依存度」と「語の吸引力」の概念を基にして、意味のある単位に文書を分解する手法を確立し、再構成するための文脈自身を探索しながら情報にアクセスすることを可能とする、ナラティブ連想情報アクセス・フレームワークの確立と実装を行った。

研究成果の概要（英文）：

This research provides users with 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. The basis of theses techniques is the notions of term dependency and term attractiveness. These notions also produced some 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. They bring out the variety of understanding and interpretation of the documents based on the Narrative Navigator framework.

交付決定額

（金額単位：円）

	直接経費	間接経費	合 計
2006 年度	4,900,000	1,470,000	6,370,000
2007 年度	4,200,000	1,260,000	5,460,000
2008 年度	3,100,000	930,000	4,030,000
2009 年度	3,000,000	900,000	3,900,000
年度			
総 計	15,200,000	4,560,000	19,760,000

研究分野：情報学
科研費の分科・細目：知能情報学
キーワード：物語、連想検索、分節

１．研究開始当初の背景

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た。従来の情報検索やデータマイニングの技術は、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た。このため、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であると考え研究を開始した。

２．研究の目的

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。

３．研究の方法

状況に応じて、動的に物語を生成する技術を実現するために、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・フレームワークについて、以下の項目に沿って研究を進めた。

① 文書における物語構造モデルの定義と、ナラティブ連想情報アクセス・フレームワークの設計
② 語の依存性に基づく語彙連鎖構造の解析
③ 文書分割方法の研究・開発
④ 文脈生成方法の研究・開発
⑤ 応用研究

４．研究成果

本研究では、物語構造に基づき、新しい文脈を生成しながら、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ついて研究・開発を行った。

(1) ナラティブ連想情報アクセス

G.ジュネットは、物語は、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)の 3 つの側面を持つとしている。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。
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 に示す。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。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。次に、ひとまとまりの出来事(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、ひとつの文書に相当する。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。

world model (世界構造)	set of stories (文書集合)
story (物語)	sequence of scenes (文書)
scene (場面)	chunk of event (段落)
event (出来事)	set of terms (文)
character (登場人物)	term (語)

表 1　物語構造モデルと文書の構成要素

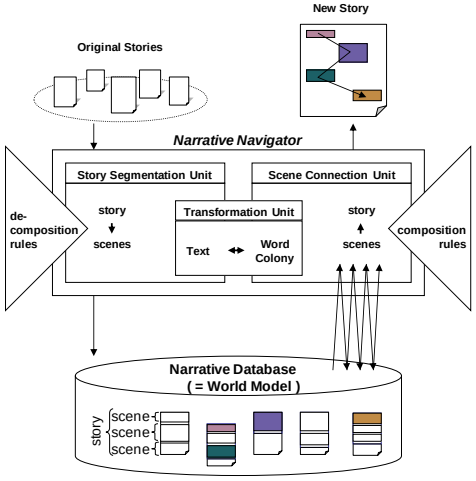

図 1　Narrative Navigator

この時、scene の抽出方法により、異なる分節が可能となり、異なる文脈に基づく story が形成される。本研究は、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、結合についての研究・開発を行った。
図 1 に、ナラティブ連想情報アクセスを可能とする Narrative Navigator (NaNa)を示す。NaNa は、物語(story)を場面(scene)に分解する Story Segmentation Unit (SSU)と、場面(scene)を連結し物語(story)を形成する Scene Connection Unit (SCU)から構成される。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・再構成機能を実現する。

(2) 語の依存性に基づく語彙連鎖構造

本研究では、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・再構成を実現する。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。

(2-1) 語の出現依存度と吸引力

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 ∈T と t'∈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')=sentences(t,t')/sentences(t),

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,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) =Σ{td(t'',t)},
(ただし、t''∈T 、t ≠ t'')

(2-2) 語彙連鎖に基づく主題俯瞰

Word Colony 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。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が考えられる。Word Colony は、(i)のグループの語を同一ノードにまとめ、(ii)の関係をノード間のリンクとして視覚化している。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。これに対して、文書の語の共起依存関係を視覚化した Word Colony は、視覚的要約と捉える事ができる。語の共起性を視覚化した他のツールでは、基本となる語の共起関係に方向性はなく、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、これにより語彙連鎖の方向性を考慮した物語生成を可能とする。

(3) 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。これを抽出するためには、文書を scene に分割する必要がある。以下に、本研究において研究・開発した文書分割法について述べる。

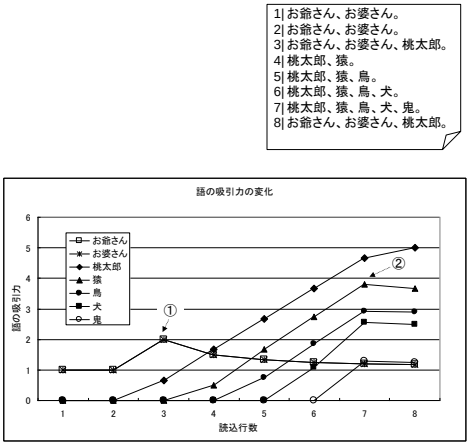

図 2　系列的文書分割

(3-1) 主題遷移解析に基づく系列的分割法

「主題遷移解析に基づく系列的分割法」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。

図 2 に例文と、読込行数に対する各語の吸引力の変化をグラフで示す。読込行数を変化させることにより、語の吸引力は、停滞部（語の吸引力が変化しないか、減少する部分）と上昇部（増加する部分）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であり、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。この例においては、"お爺さん"に着目した場合、 3 行目が"お爺さん"で表されるトピックの終了であり、"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。

この分割方法は、吸引力の大きなメイン・トピックだけに着目して、文書分割をするのではなく、吸引力は小さくても、ユーザが着目しているトピックに関しての文書分割が可能であり、分割によって、ユーザの着目するトピックがメイン・トピックとなる物語生成が可能となる。

(3-2) 主題階層解析に基づく場面的分割法

「主題階層解析に基づく場面的分割法」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 を抽出する手法である。

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

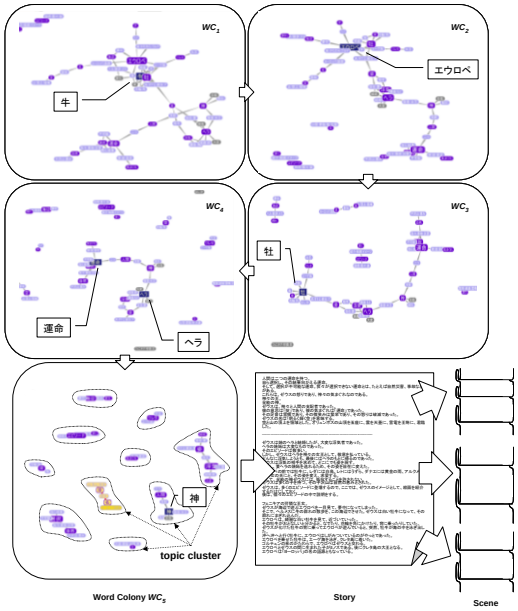

図 3　場面的文書分割

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。これは、ある主題に関して、異なる観点で集められた情報として解釈できる。

(4) 語彙連鎖に基づく文脈生成

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。

本研究においては、各種の文書集合に対して、以下に述べるようなトピック遷移パターンの解析を行い、応用分野において適切な物語生成を支援するシステム構築を目指す。

図 4 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

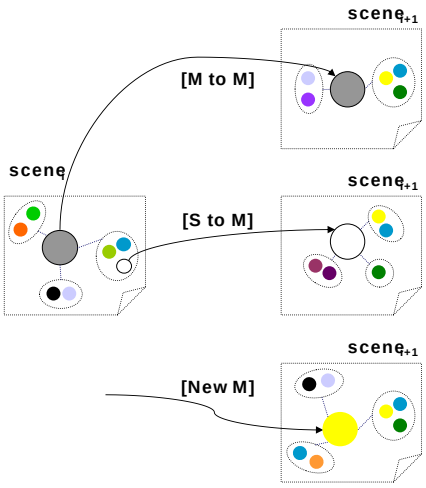

図 4　トピック遷移パターン

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。

図 5 のグラフは、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。

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 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。

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。

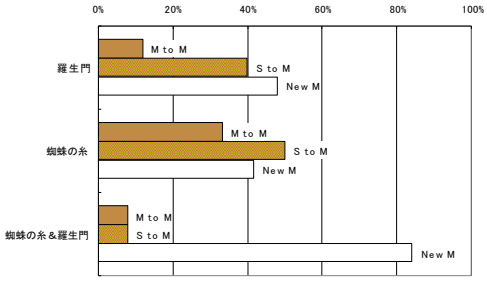

図 5　トピック遷移パターンの出現割合

(5) 応用に関して

ナラティブ連想情報アクセスにおいて、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。

右図は、小型衛星の設計議事録に対する、ナラティブ連想情報アクセスの例を示す。図中には、ユーザの要求に応じて、連鎖情報をもとに生成された複数の文脈が示されている。これらを横断して辿る行為自体により、新しい知識生成を促すことが可能となる。小型衛星の開発においては、最先端の技術が多用されており、発生するトラブルの原因が文書化されていないことは多々生じ、通常の検索エンジンでは、原因究明にはたどり着けなかった。図に示した事例研究においては、小型衛星の運用時に起きた実際の機器故障トラブルから、その原因へつながる文脈を設計議事録集合の中からナラティブ連想情報アクセスにより発見できることを示した。

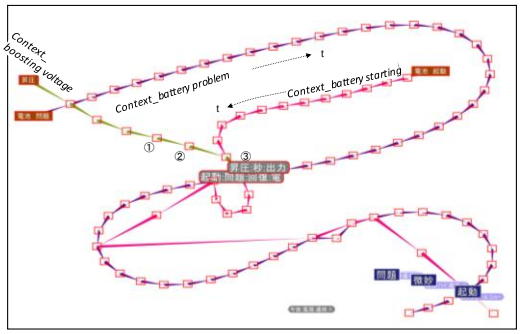

図 6　ナラティブ連想情報アクセス

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　11 件）

① 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年6月)

② Mina Akaishi, Nicolas Spyratos, Koichi Hori and Yuzuru Tanaka: Connecting Keywords Through Pointer Paths over the Web, LNCS, Vol.3847, pp.115-129, 2006

③ Mina AKAISHI, Taizo YAMADA, Tetsuya ISHIKAWA and Koichi HORI: 13th International Conference on Information Visualization (IV09), pp.572-576, 2009

〔学会発表〕(計 21 件)

① 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, 人工知能学会人工知能基本問題研究会資料, SIG-FPAI-A603, pp.29-34, 2007 年 3 月

② 加藤義清, 赤石美奈, 堀浩一, 時間属性付き文書集合からの潜在多重文脈の抽出、人工知能学会人工知能基本問題研究会研究会資料, SIG-FPAI-A603, pp.41-44, 2007 年 3 月

③ 大東 誠: 実世界の空間情報に基づく情報サービスの統合・検索のための位置モデル, 電子情報通信学会知能ソフトウェア工学研究会(2007 年 7 月)

〔図書〕(計 0 件)

〔産業財産権〕
○出願状況(計 0 件)

○取得状況(計 0 件)

〔その他〕
ホームページ等

６．研究組織

(1)研究代表者
赤石 美奈 (AKAISHI MINA)
東京大学・先端科学技術研究センター・准教授
研究者番号：60273166

(2)研究分担者
堀 浩一 (HORI KOICHI)
東京大学・先端科学技術研究センター・教授
研究者番号：40173611
(平成 18 年度、平成 19 年度)

田中 克明 (TANAKA KATSUAKI)
東京大学・先端科学技術研究センター・助教
研究者番号：80376657
(平成 18 年度、平成 19 年度)

加藤 義清 (KATO YOSHIKIYO)
独立行政法人情報通信研究機構・第二研究部門知識創成コミュニケーション研究センター・研究員
研究者番号：50373444
(平成 19 年度)

大東 誠 (OHIGASHI MAKOTO)
東京大学･大学院情報学環･助教
研究者番号：40421995
(平成 19 年度)

(3)連携研究者

堀 浩一 (HORI KOICHI)
東京大学・先端科学技術研究センター・教授
研究者番号：40173611
(平成 20 年度、平成 21 年度)

田中 克明 (TANAKA KATSUAKI)
東京大学・先端科学技術研究センター・助教
研究者番号：80376657
(平成 20 年度、平成 21 年度)

加藤 義清 (KATO YOSHIKIYO)
独立行政法人情報通信研究機構・第二研究部門知識創成コミュニケーション研究センター・研究員
研究者番号：50373444
(平成 20 年度、平成 21 年度)

大東 誠 (OHIGASHI MAKOTO)
東京大学･大学院情報学環･助教
研究者番号：40421995
(平成 20 年度、平成 21 年度)