社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－NL－181 (12)
2007／9／26

テキストの話の流れを視覚化するインタフェース–川下りシステム–

砂山 渡†

コンピュータとインターネットの普及に伴い，電子テキストを作成，閲覧する機会が増えている．多くのテキストの内容を素早く理解するためには，ひとつひとつのテキストの主題，そしてその話の流れを素早くつかむことができる環境が望まれる．自動要約システムを用いることでもテキストの概要を知ることができるが，指示的要約では，のちに実際のテキスト本文を参照する必要がある．また，報知的要約では，一定量のテキストを実際に読む必要が生じるため，より簡潔にテキスト全体の話をつかめる環境があると便利である．そこで本研究では，テキスト中の各単語に，テキストの主題との関係を表すラベルを付与し，どのような単語が，テキスト中のどのくらいの位置で出現し，各単語や各段落が主題とどのような関係にあるかを明確にする視覚化インタフェースを提案する．

Text Stream Visualization Interface –River Rafting System–

WATARU SUNAYAMA†

We have many occasions to read electrical texts along with the growth of computers and the internet. An environment to grasp whole contents and flows is required when we comprehend those texts quickly. Automatic summarization methods are also used for this purpose but indicative summaries require our reading whole texts after using them. Since informative summaries also have some quantity, the simple architecture to know the whole texts become useful. In this study, the system labels each word in a text according to the theme, and visualizes labeled words on the interface to know when and how each word is related to the theme.

1. はじめに

コンピュータとインターネットの普及に伴い，電子テキストを作成，閲覧する機会は増加の一途を辿っている．検索システムやデータベースシステムの発展により，電子テキストを獲得することは容易になってきたが，人間がそれらのテキストを読んで処理する能力が変わるわけではないため，得られたテキストの取捨選択や，読むべきテキストを素早く閲覧，理解するための環境が望まれるようになってきた．

得られたテキストが真に有効であるか否かの判定にはベクトル空間モデル[2]などに基づくキーワードベースのマッチングや，人手や自動要約システムによる要約を参考にすることが多い．

検索結果によって得られたテキストに対しては，検索エンジンによる指示的要約が与えられることが多いが，指示的要約は興味があるテキストへのポインタとなることが主な役割であり，テキストの内容全体を俯瞰する目的で使用するには不十分と考えられる．また，テキストの内容全体を知るための報知的要約は一般に情報の圧縮率が低く，もとのテキストの少なくとも30%以上を読む必要が生じる[7]ため，テキスト全体の内容を素早く知るという目的において，より効果的な環境が望まれる．

本研究においては，テキストの内容を文ではなく単語で表現する．テキストからキーワードを抽出する手法や，テキスト中の単語間の関係を視覚化する研究は多数存在する．しかし，各単語がテキストの話の流れの中の，どの辺りで出現して，どこでテキストの主題との関係を持ったのか，また，テキスト全体として主題に関する一貫性がどの程度存在するのか，というテキストの話の流れに着目して視覚化した研究は少ない．

本研究では，テキスト中の各単語に，テキストの主題との関係を表すラベルを付与し，この主題との関係を明確にする視覚化インタフェースを提案する．本稿で提案するラベル付け手法は，議論支援を目的とするラベル付け手法[6]を改良したものであり，このラベル付けがテキスト分析に一定の効果があることは確認されているが，ラベル付けの結果を容易かつ直感的に解釈できる環境が存在していない．そこで，ラベル付けされた各単語，および各ラベルが与えられる単語の総数に着目し，テキスト全体として，主題に関係する

† 広島市立大学大学院 情報科学研究科，Graduate School of Information Sciences, Hiroshima City University
sunayama@sys.im.hiroshima-cu.ac.jp

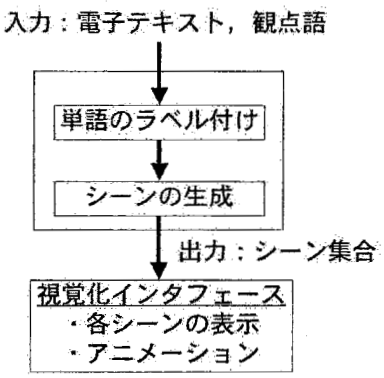

図 1　川下りシステムの構成

単語がどこでどの程度使われているかを把握でき，主題に関する一貫性の存在の有無を確認できるインタフェースを提案する．

2. 川下りシステム

本章では，話の流れの理解を支援するための川下りシステム（図 1）について説明する．

川下りシステムでは，電子テキストとその主題を表す単語集合（観点語と呼ぶ）を入力として，テキスト中で使われている単語に，テキストの主題との関係を表すラベルを付与した上で，テキストの途中位置までにおけるラベル付けの状況を表すシーンを生成する．それら各シーンの集合をもとに，視覚化インタフェース上で，特定のシーンの表示したり，シーンを連続的に変化させるアニメーションを再生することにより，ユーザがテキストの話の流れを理解することを助ける．

2.1 テキストの流れと川下りとのアナロジー

テキストにおいて意味の関連がある単語が連続的に出現するという語彙的連鎖[1]の性質を踏まえた，「流れが明確なテキストにおいては，単語が連続的に出現する」というサブトピックモデル[4]により，テキストに出現する単語を水に，テキストの話の流れを川の流れに喩える．

テキストの話が始まった時点においては，使われた単語の量も少なく，またどの単語も新しく出現する単語になる．これらの単語が川の最も上流を流れる単語となり，そこからこの本流の流れが始まっていく．

無事に，この流れが海に到達して話が完結するためには，より多くの単語の投入が必要となってくるが，周りを流れる水源の水すべてが，この本流に注ぎ込み，話の展開を助けるわけではない．どんな水でも本流に汲み入れていると，かえって水があふれてしまい，目的とは違う方向に流れが進んで行ってしまうことになる．

話がある程度展開していき，本流の水量が増してくると，本流から分岐する支流が現れるようになる．支流は，テキストの主題に関係する内容であるが，本流の話の流れに比較して，詳細で具体的な話が展開していると考えることができる．やがて，支流は他の水源からの水を得つつ，再び本流と合流することもあるし，そのまま別の方向に流れていくこともあるだろう．

最終的に，本流が十分に太くなれば，話が十分に展開されたと考えることができ，話は収束を迎えることになる．本稿で提案する視覚化インタフェースでは，テキストの話の流れをこのような川の流れとして表すことで，直感的な話の流れの理解を支援する．

2.2 電子テキストと観点語の入力

本稿で扱うテキストは，複数のセグメント（段落などのテキスト内での意味の区切り）から構成されており，各セグメントは単一もしくは複数の文から構成されているとする．また，各セグメントや文の区切りは特定できると仮定する．観点語は，テキストに含まれている単語のいずれかを，ユーザ自身が与えることもできるが，ユーザが観点語を与えない場合においても，自動要約システムのひとつである展望台システム[5]によって自動的に抽出して与えられる．展望台システムは，一文内での単語の共起性をもとに，多くの単語と同時に現れる単語を観点語として抽出する．すなわち，テキスト全体を通じて出現し，最も一貫性がある単語を抽出するため，テキストの話の流れを表す本システムの観点語として適切である．

2.3 単語間の距離とテキストの主題との関連

テキスト中の各単語にラベル付けを行なうための準備として，単語間の類似度を表す単語間距離を定義する．まず，セグメント Seg 内における単語 w_i, w_j 間の距離（セグメント内距離）を式 (1) で定義する．

$$distance(w_i, w_j, Seg) = \min_{s \in Seg} |ap(w_i, s) - ap(w_j, s)| \quad (1)$$

ただし $ap(w, s)$ は，単語 w が文 s 内で出現した箇所が，文の先頭から何単語目であるかを与える関数とする．すなわちセグメント内距離を，セグメント内の各文 s において単語 w_i と単語 w_j が出現した箇所間の距離（w_i, w_j 間の単語数）の最小値☆として定める．

単語間の距離を，シソーラスなどによって与えない理由は，同じ単語同士でもテキストが異なれば，単語間の類似性も異なると考えたためである．

各テキストには，そのテキストで述べたい主題が存在する．本稿では，後述する単語のラベル付けの際に，各単語が主題に関係しているか否かを重視する．そこで，本節で定義したセグメント内距離を用いる．すなわち，各セグメントにおいて，テキストの主題を表す観点語の集合 T に含まれるいずれかの単語とセグメ

☆ 単語 w_i と単語 w_j の片方でも出現しないときは無限大とする．

ント内距離が，しきい値 $dmax$ 以下の単語を主題との直接の関係が理解できる「TOPIC 関連語」，しきい値 $dmax$ より大きい単語を主題との関係が明らかでない「TOPIC 非関連語」とする．また現在，$dmax$ は 10 としている☆．

結果として，観点語と同一文中に現れ，かつ観点語と $dmax$ 単語以内に出現する単語が主題に関係する単語となる．またセグメント内で複数回出現する単語については，セグメントの後半で主題に関係する単語となる場合であっても，セグメントの前半部分においても主題に関係する単語になる．これは，セグメントの大きさにも依存するが，セグメントという狭い範囲内では，同じ単語は同じ文脈で使われて単語間の関係は維持されると仮定するとともに，「定義 (主題との関連づけ)」と「使用」が前後になることをセグメント内でのみ許容することに相当する．

2.4 単語のラベル付け

テキスト中の各単語☆☆ には，テキストの主題との関係を表すラベルを与える．

以下に，テキスト内の各単語に与えられる 6 つのラベルとその意味を示す．

- TOPIC：テキスト全体の観点となる単語．観点語．人手により与えるか，展望台システムで自動抽出する．意味：テキストの主題を表す単語．
- SEED：観点語の近くに，初めて現れた単語．意味：主題に関する話を広げる際に，話の種として用いられる単語．
- MAIN：観点語の近くに現れて，主題に関する論理を展開する単語．意味：話の本筋に強く関わる単語．主題を構成する単語．
- NEW：観点語から離れた場所に，初めて現れた単語．意味：1) 主題の一部分を深く掘り下げた話 (副主題) を広げる際に用いられる単語．2) 主題と関係のない話に用いられる単語．
- BYWAY：一度も観点語の近くに現れないまま，複数回使用されている単語．意味：1) 副主題にのみ関係して，繰り返し使われている単語．冗長な単語の可能性がある．2) 主題と関係のない話に繰り返し用いられている単語．
- SUB：観点語の近くに現れた後，観点語から離れて現れた単語．意味：主題に関する話に用いられた後，副主題として展開した話にも関係する単語．話の副主題を構成する単語．

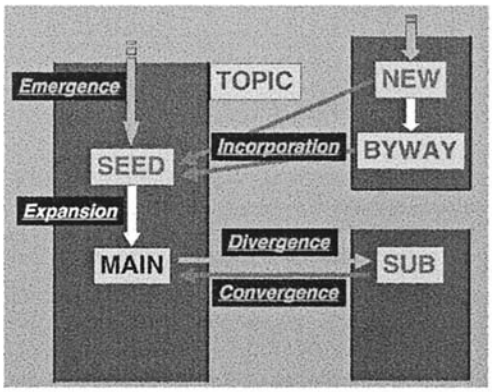

図 2 同一単語に与えられるラベルの遷移パターン

ラベル付けは，テキストの前から 1 セグメントごとに以下の手順で行なう．

1. 単語が観点語であれば TOPIC のラベルを与える．
2. 単語間の距離を表す式 (1) によって，セグメント内の各単語が，観点語と距離 $dmax$ 以内の TOPIC 関連語であるか否かを調べる．
3. テキスト中で初めて出現した単語が TOPIC 関連語であれば SEED，そうでなければ NEW のラベルを与える．
4. テキスト中で 2 回目以上の出現となる単語が TOPIC 関連語で，今までに一度も TOPIC 関連語になったことがない単語には SEED，過去に TOPIC 関連語になったことがある場合には MAIN のラベルを与える．
5. テキスト中で 2 回目以上の出現となる単語が TOPIC 非関連語で，今までに一度も TOPIC 関連語になったことがない単語には BYWAY，過去に TOPIC 関連語になったことがある場合には SUB のラベルを与える．

2.5 単語ラベル遷移アーク

同一単語の与えられるラベルが，各セグメント間でどのように遷移するかを図 2 に示す．この遷移関係をアークと呼び，ラベル間の遷移を表す 7 つのアークとその意味を以下に示す．

- Emergence (未出現→ SEED)：主題に沿った新たな単語が現れること．
- Expansion (SEED → MAIN)：主題に沿って一度しか現れていなかった単語を用いて，話を膨らませた．
- Incorporation (NEW or BYWAY → SEED)：観点語から離れて現れていた単語が，初めて観点語の近くに現れること．主題との関係が明確でなかった単語を，初めて観点語を共に用いて話をし

☆ これは，意味のある一文に含まれる自立語の数が最低 7 単語である[5] ことと，その最も短い意味のある文が並んだときに，主題に関連する文とその前後の文を関係がある範囲の上限とみなし，合計 21 単語の中心に観点語があると仮定して，$(21-1)/2 = 10$ という計算により定めた．

☆☆ なお以下の本稿では「単語」として，形態素解析器[8] を用いて抜き出される「名詞」を対象としている．これは，一般的にテキストに与えられるキーワードが名詞であることと，単語の認識と理解のしやすさを重視したことによる．ただし，目的に応じて「動詞」「形容詞」など他の品詞を追加することは可能である．また同様の理由で，本稿の実験時には，一文字の単語および平仮名で始まる単語は除いた．

たことにより，主題との関係を明らかにした．一見関係のない話や，とりとめのない話の中から，話題のヒントが現れた状況にも相当する．

- Divergence（SEED or MAIN → SUB）：観点語の近くに現れていた単語が，観点語から離れて現れること．観点語の出現頻度を相対的に下げて，主題に関係するある単語について，深く掘り下げた話を始めること．話の発散．
- Convergence（SUB → MAIN）：観点語から離れて現れていた単語が，再び観点語の近くに現れること．一時的に深く掘り下げた話をしていたため観点語の近くに現れなかった単語を，再び観点語と結びつけて話をしたことにより，掘り下げた話の結論と主題との関係を明らかにした．話の収束．
- Out-Emergence（未出現→ NEW）：主題との関係が不明な新たな単語が現れること．
- Out-Expansion（NEW → BYWAY）：主題との関係が不明な一度しか現れていなかった単語を用いて，主題との関係が不明な話を膨らませた．

これらのラベル遷移アークは，次章で述べる視覚化インタフェース上で，ユーザがアニメーションを見て，テキストの話の流れ，特に各セグメントの話の流れの中での位置付けを理解するために用いられる．

3. 視覚化インタフェース

本章では，前章で定義した，テキスト中の各単語に与えられるラベルと，ラベルの遷移パターンに与えられるアークとを直感的に理解できる視覚化インタフェースについて述べる．すなわち，図2と同じ形式の二次元インタフェースを図3のように構築した．

インタフェースは，左の川と右上の水源，右下の川と，6つのボックス，ボックス間をつなぐ7本の線から構成されている．6つのボックスと7本の線は，それぞれ，単語のラベルと，単語の遷移を表すアークとに対応しており，真ん中上部のボックスにTOPIC，左の川にあるボックスの上がSEEDで下がMAIN，右上の水源内のボックスの上がNEWで下がBYWAY，右下の川のボックスがSUBのそれぞれのラベルが与えられた単語が入れられる．すなわち，左側の川がテキストの主題に関係する本流(SEED,MAINを含む)を表し，右上に主題に未関連な水源を，右下の川が本流と平行して流れる支流を表している．

インタフェース上に表示する単語は，テキストの最初から任意の途中のセグメントまでに出現した全単語であり，指定するセグメント終了時点で各単語に与えられているラベルに対応するボックスに，各単語が収められる．

またこのインタフェースでは，各セグメント終了時点の状態（各単語のラベルとその位置）を1シーンとして，連続するシーンを滑らかに補間する（単語がボックス間を移動する）アニメーションを再生することができる．これによって，テキストの最初から最後までの単語の出現状況や，ラベルの変化を続けて見ることができる．

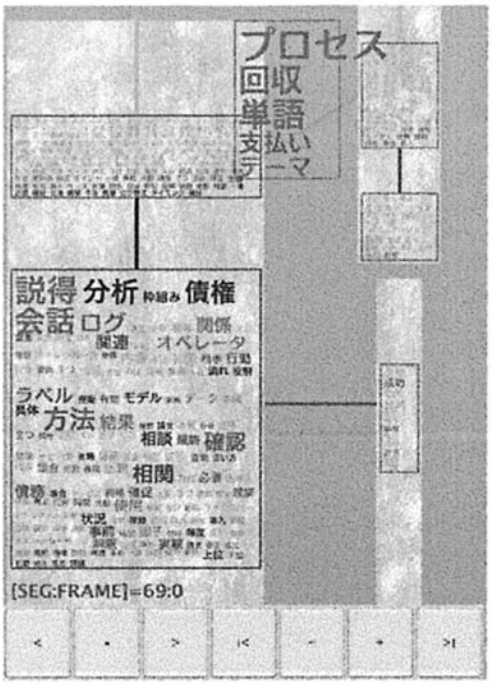

図3 川下りシステムインタフェース

また，単語の流れの理解を助けるため，以下の工夫を施している．

1. 各単語は各ボックス内で，テキスト内での単語の出現順に表示される．
2. 各単語は，そのラベルに対応した色がつけられる．ただし，NEW,BYWAYを経由した単語は，NEW,MAINのボックス内ではその色を維持する．一度SUBになった単語は，以降ずっとSUBの色を維持する．
3. 単語の出現頻度が大きくなるにつれ，フォントサイズが大きくなる．
4. アニメーション時に，ボックス間を移動する単語数に応じてアークが太くなる．
5. 過去10セグメント現れなかった単語は，徐々に薄くなっていく．

1.はテキストのどの辺りで出てきた単語なのか，また単語の出現順序関係の情報を与えるためである．2.のNEW,BYWAYの色に関しては，もともと主題に関係する単語だったのか，後から主題との関係を与えられた単語なのかを区別するため，SUBの色は一度深く掘り下げられた話題に関係する単語であることを

表 1 実験に用いたテキスト

ジャンル	TEXT	SEG	LABEL	LABEL /SEG
論文の原稿	5	59	1818	31
昔話	5	31	191	6
ニュース記事	10	10	165	16
コラム記事	5	9	118	13
ブログエントリ	11	20	149	8
ブログコメント	5	100	954	10
掲示板	8	74	691	9

明示するために設けた．3. は単語の出現頻度による重要度を表すためであり，4. は各セグメント間での単語の移動を要約する目的で設けた．5. は全ての単語を表示することで煩雑になることを避けるとともに，各セグメント終了時点におけるテキストの読み手の思考状態に近いイメージを作成することを意図している．

4. ラベル付けアルゴリズムの評価

テキスト中の各単語に与えられるラベル，およびラベルの変化を表すアークの妥当性を評価する実験を行なった．用意したテキストは，表 1 に示す，論文の原稿，昔話，Web 上のニュース記事とコラム記事，ブログ (日記)，ブログコメント，掲示板の計 49 テキストである．いずれも，テキストの流れを把握することを目的としているため，極端に短いテキストは含んでいない．また，ブログコメントと掲示板においては，一部の人々の間で話題になるなど，盛んに書き込みが行われていたブログや掲示板の中から，最初の 100 コメントを抜き出して用意した．ただし，表中の TEXT は各ジャンルのテキスト数，SEG はセグメント数，LABEL はラベルが付与された単語の総数，LABEL/SEG は 1 セグメント当たりのラベルが付与された単語数を表し，各数値は各ジャンル内での平均となっている．またセグメントの与え方は，ブログコメントと掲示板のテキストに関しては，1 つのコメントや書き込みを 1 つのセグメントとし，それ以外のテキストでは，空行や段落の切れ目をセグメントの区切りとした．

表 2 に，テキスト中の単語に，各ラベルが与えられた割合（各ジャンル内での平均値）を示す．ただし，TOPIC 関連は，TOPIC,SEED,MAIN,SUB の 4 つのラベルの合計を，TOPIC 未関連は，NEW,BYWAY の 2 つのラベルの合計値を表す．

論文の原稿では，MAIN の割合が高く，主題にそった多くの説明がなされていたことが伺える．昔話では，TOPIC の値が高い反面，MAIN の割合が低くなっており，物語の主人公がさまざまに場面が変化する中で話が展開したことに対応すると考えられる．ニュースやコラム記事では，主題に関係する新しい単語である SEED の割合が高く，幅広い情報提供の意味合いが強かったことが伺える．ブログエントリでは，MAIN の割合が低く，主題に沿った一貫した話ではなく，ブログの作者が思いつきのままに，さまざまなことを書いていたと考えられる．ブログコメントと掲示板では，BYWAY の割合が高く，主題とは関係のない話題についての話が続きやすかったことが伺える．その他，ブログエントリ，昔話，掲示板の順に主題に未関連の新出単語 NEW の割合が高くなった理由は，これらのテキストでは，話の流れが読めず，全く予想外の方向に話が進むことが多かったためと考えられる．また，ブログコメントや掲示板において，単語が関連と非関連の間を行き来するラベル SUB の割合が高くなっており，これは対象としたテキストにおいて，類似の議論が繰り返し起こり議論のループが見られたことが原因と考えられる．

表 2 における，TOPIC 関連のラベルを含む割合は，論文の原稿，ニュースとコラム，ブログコメントと掲示板と昔話，そしてブログエントリ，の順になっており，これは実際のテキストの内容が，どの程度主題に関係する話を多く含んでいたかを反映した結果になっている．すなわち，論文の原稿には主題に関係する単語が繰り返し出現するため，主題との関係が強いと判断され，ニュースやコラム記事については，主題に関係して短く完結にまとめられている．昔話は主題に関する一貫性があると予想される反面，さまざまな伏線や背景描写の記述も含まれるため，全体としての TOPIC 関連度は下がる．掲示板やブログのコメント欄については，その掲示板の主題や対象となるブログエントリの制約の下で，比較的自由な記述が可能であったため，TOPIC との一定の関連度を保ちつつも，特に高い値にはなっていない．ブログエントリは，書き手が何の制約も受けずに記述することが可能なため，最も主題に関する揺れが生じたと考えられる．以上のことから，システムのラベルづけによって表される各テキストの主題との関連度は妥当な数値だったと言える．

表 3 に，テキスト中の単語のラベル変化を表すアークの割合（各ジャンル内での平均値）を示す．ただし，各アークの表記は，2.5 節で定義した表記の頭文字部分を用いている．また，表中の ARC はアークの総数，Conv/Div はアーク Div の数に対するアーク Conv の割合で，SUB のラベルが与えられた後に再び MAIN のラベルが付与された単語の割合を表し，ARC/LABEL はラベルが与えられた単語総数に対するアーク総数の割合で，同一単語が再び出現する際にラベルの変化が起こった割合を表す．

アーク Em とアーク O-Em は単語の出現を表し，ラベルの SEED と NEW とほぼ同じ意味であるため，SEED が多かったニュース記事とコラム記事において Em の割合が，NEW が多かったブログエントリ，昔話，掲示板において O-Em の割合が高くなった．SUB のラベルが多かったブログコメントと掲示板テキストにおいて，アーク Div の割合が高い反面アーク Conv の割合が高くなく，話が発散的であったことが伺える．

表 2　テキスト中の単語に各ラベルが与えられた割合 (%)

ジャンル	TOPIC	SEED	MAIN	NEW	BYWAY	SUB	TOPIC 関連	TOPIC 未関連
論文の原稿	22	15	51	6	2	4	92	8
昔話	28	23	8	31	7	4	63	37
ニュース記事	20	41	13	19	2	4	78	21
コラム記事	20	41	12	21	2	4	76	24
ブログエントリ	19	30	8	32	6	5	55	46
ブログコメント	17	23	17	24	8	11	68	32
掲示板	17	22	14	27	12	8	61	39

表 3　テキスト中の全アークに対するアークが与えられた割合 (%)

種類	ARC	Em	Exp	O-Em	O-Exp	Inc	Div	Conv	Conv/Div	ARC/LABEL
論文の原稿	609	38	23	16	4	5	9	6	71	34
昔話	125	31	8	38	7	7	7	2	34	65
ニュース記事	123	52	11	27	3	2	4	1	16	74
コラム記事	88	54	9	27	3	2	5	1	15	74
ブログエントリ	112	33	6	41	7	4	7	1	17	75
ブログコメント	618	32	10	32	6	5	11	4	37	65
掲示板	449	26	9	38	9	5	9	4	42	65

表 4　単語出現総数とラベル出現割合との相関係数

TOPIC	SEED	MAIN	NEW	BYWAY	SUB
-0.11	-0.51	0.83	-0.45	-0.02	0.36

表 5　単語出現総数とアーク出現割合との相関係数

Em	Exp	O-Em	O-Exp	Inc	Div	Conv
-0.47	0.15	-0.45	-0.17	-0.01	0.25	0.66

論文の原稿においてもアーク Div の割合が高かったが，同時に Conv の割合も高く，一方的に話を広げて発散させるだけではなく話をまとめようとしていたためと考えられる．アーク Inc の割合は，もともとラベル NEW の数が多かった昔話や掲示板テキストで高くなったことに加え，論文テキストでは NEW の割合が最も低かったにもかかわらずそれらのテキストと同程度の割合があった．このことから，積極的に主題との関係を明確にする意図があるテキストにおいては，アーク Inc の割合も高くなると考えられる．

表 4 に，全 49 テキストの単語出現総数と各テキストのラベル出現割合との相関係数を，表 5 に，全 49 テキストの単語出現総数と各テキストのアーク出現割合との相関係数を示す．表 4 の SEED,NEW のラベルおよび表 5 の Em,O-Em のアークの出現割合に中程度の負の相関が出ているように，テキストが長くなると，新しい単語が現れにくくなっている傾向が確認できる．逆に，ラベル MAIN,SUB との間に正の相関が表れており，テキストが長くなるにつれて，既に使った単語が使われるようになっている．また，ラベル SUB との正の相関と関係して，アーク Div,Conv との間に正の相関が表れているが，特にアーク Conv との相関値が高くなっている．アーク Conv が現れるためには，MAIN のラベルが与えられた単語がアーク Div を通じて，ラベル SUB が与えられている必要があり，アーク Div で表される発散が，後にアーク Conv で収束する割合 (表 3 の Div/Conv) は，論文テキストで 71%となっている以外は，15%から 42%と低い割合にとどまっている．長いテキストは，論文の原稿，ブログコメント，および掲示板のテキストであることから，テキストが長くなるにつれて，話の収束が起こりやすくなった，もしくは同じ話の繰り返しが起こりやすくなったと解釈することができるとともに，話の収束を起こすためには，話の長さも必要な要素のひとつであると考えることができる．

表 6 に，ある論文の原稿テキストおよび昔話テキストにおいて，セグメントの区切りを変化させてラベル付けを行なったときに，与えられるラベルがどのように変化するかを調べた結果を示す．セグメントの大きさを変えることによって現れるラベルの変化は，ラベル付けのアルゴリズムにおいて，セグメント内のある単語が「TOPIC 関連」もしくは「TOPIC 非関連」で統一されるという点となる．このことから，セグメントが大きくなるほど，ある単語が「TOPIC 関連」があると判定される可能性が高くなる．そのため，表 6 の結果にも，MAIN の割合がセグメントが小さくなるにつれて減少し，NEW や SUB の割合が増加することにつながって現れている．直感的には，SEED の割合も減少し，BYWAY の値が増加することが予想される．しかし SEED はテキスト中で各単語が一度だけ与えられるラベルであり，テキストの主題を表す観点語と同じ文中に現れることが「TOPIC 関連」と判定されるための必要条件であるため，その一度はセグメントの長さには依存しない．BYWAY は，表の割合には現れないごくわずかな数で変化が見られたが，テキストを通じて未関連のままの単語が多かったこと，ま

表 6　テキストをセグメントに区切る箇所の違いと単語ラベル付与割合 (%) との関係

テキスト	SEG の区切り	SEG	LABEL /SEG	TOPIC	SEED	MAIN	NEW	BYWAY	SUB	TOPIC 関連
論文 A	章	7	265	21	16	56	5	2	1	93
論文 A	章と節	22	84	21	16	54	5	2	3	93
論文 A	章と節と段落	69	27	21	16	52	6	2	4	92
昔話 A	大段落	4	60	32	26	14	19	6	3	75
昔話 A	文	94	3	32	26	10	20	6	6	74

た未関連から関連にラベルを変えることを表す Inc のアークが少なく，図 2 で NEW から BYWAY を経ずに SEED となる単語が多かったため，大きな変化が現れなかった．したがってセグメントの大きさは，テキスト中で「TOPIC 関連」と「TOPIC 非関連」の関係を行き来する MAIN と SUB のラベルに，最も大きな影響を及ぼすことが分かる．SUB のラベルが与えられる単語は，テキストの流れの中で特に話が掘り下げられている話題に関係する単語であると考えられるため，テキストの流れの細かい変化を捉えたいときには，セグメントの大きさを小さめにし，テキストのおおまかな流れを捉えたいときには，セグメントの大きさを大きく設定するのがよいと考えられる．

5. 視覚化インタフェースの使用例

本章では，テキスト中の各単語に与えられるラベルおよびアークをもとに，テキストの話の流れを直感的に把握するために，視覚化インタフェースを用いた例を示す．

テキスト中の各単語に与えられるラベルの量が，視覚化インタフェース上の川幅を表現し，アークがラベルの変化、すなわち川の流れの変化としてアニメーションで表される．すなわち，視覚化インタフェースにおける見るべきポイントとその意味は下記で表される．

1. 本流の川幅：テキストの主題に関係する単語の数
2. 支流の川幅：テキストの主題に関係する話の中で，特に掘り下げられた話に関係する単語の数
3. 水源の水量：テキストの主題に未関連な話のもとになりえる単語の数
4. 水源の水量や川幅の変化を与えるタイミング：テキストの話の流れに影響を与えるセグメント
5. 表示される単語の濃さや大きさ：テキスト中で重要な位置付けにある単語

図 3 に論文テキストを用いた時の最終セグメントにおけるインタフェース表示を，図 4 に掲示板テキストを用いたときの最終セグメントにおけるインタフェースの表示を示す．前者は「説得プロセス分析の枠組みと債権回収会話ログへの適用」というタイトルの論文であり，展望台システムによって自動抽出された観点語は「単語，プロセス，回収，テーマ，支払い」であった．また後者には，ある戦闘ゲームの立ち回りに関する議論掲示板を用い，自動抽出された観点語は「武器，

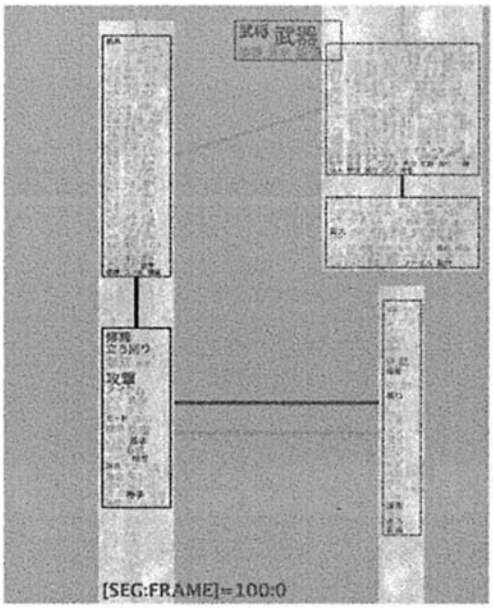

図 4　インタフェース画面：掲示板テキスト

極書，武将，装備，背水」であった．図 3 においては，左の本流が太く，多くの単語が繰り返し使用されて濃く表示されているのに対して，図 4 では，多くの単語が薄く表示されていること，本流が細いこと，右上の水源の量が多くなっていることが確認できる．このように本インタフェースを用いることで，テキストの主題に関わる単語がどの程度使われているかを直感的に把握することができ，またどの程度主題に関係する話が展開されてきたかを確認することができる．

図 5 に，論文テキストを用いてアニメーションを行なっている途中（全 69 セグメント中，38 セグメント目のテキストによる単語ラベルの変化）のインタフェース画面を示す．本流の上方の SEED ボックスの上部の線，SEED ボックスと下の MAIN ボックスをつなぐ線，および MAIN ボックスと支流の SUB ボックスをつなぐ線が，それらのボックス間を移動する単語の量に比例して太くなっており，アニメーションにおいては，それらの単語が動く様子と合わせて確認することができる．このアニメーションによって，そのセグメント内の単語が，テキストの流れに与える影響をひ

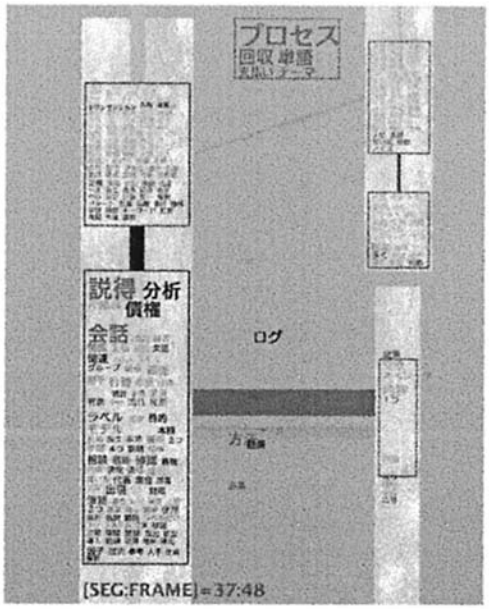

図 5　アニメーション中のインタフェース：論文テキスト

と目で確認することができる．例えば，このシーンの解釈としては，水源のまわりの単語に動きがないことから，テキストの主題に関わる話のみが展開されていること，また SEED ボックスの上下の動きから，新しい単語を用いながら今までの話を発展させた話題を展開させていること，また支流の方に単語が流れていることから，主題に関して踏み込んだ議論がなされていると予想できる．実際のテキストにおいては，「会話ログの分析方法」という小節が始まり，その方法の詳細についての説明がなされており，シーンの解釈と一致する．

以上のことから，本視覚化インタフェースは，テキストの主題との関係に基づく話の流れ，および各セグメントのテキストの主題との関わりと位置付けを直感的に理解するために有効に役立てられると考えられる．

6. 結　　論

本稿では，テキスト中の各単語に，テキストの主題との関係を表すラベルを付与し，テキストの話の流れを明らかにする川下りシステムとその視覚化インタフェースについて述べた．本システムがテキスト中の各単語に与えるラベルの妥当性を検証し，各ラベルが与えられた単語の全単語に対する割合を見ることにより，テキストが主題に関して一貫性があるかを判断できることを確認した．

本システムは，一度に読み切ることができない長いテキストや，コメントの多い電子掲示板などに対して用いることで，その全体像を把握することに役立てられると期待できる．また，論文などのテキストを作成する際に，主題に関する一貫性の有無や，各単語と主題との関わり方を確認することで，テキストの推敲支援にも用いることが可能と考えている．

今後は，本インタフェースを利用するユーザが，テキストの流れと内容を直感的に素早く理解できることを，実験により検証していく．

謝　　辞

本研究の一部は，文部科学省科学研究費補助金，基盤 A(1)「組織的チャンス発見を支援するシナリオマップ・システム」（課題番号：16200006），ならびに若手 (B)「大規模データの意味理解と解釈を支援する思考活性化インタフェース」（課題番号：19700151）の援助を受けた．

参 考 文 献

[1] Morris, J. and Hirst, G.: "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", Computational Linguistics, Vol.17, No.1, pp.21 – 48, (1991).

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

[3] Wataru Sunayama and Masahiko Yachida: Panoramic View System for Extracting Key Sentences Based on Viewpoints and an Application to a Search Engine, Journal of Network and Computer Applications, Elsevier Science, Netherlands, Vol.28, No.2, pp.115 – 127, (2005).

[4] 砂山渡，橘啓八郎：サブトピックモデルに基づく文章の流れの評価指標の提案，日本知能情報ファジィ学会誌，Vol.18, No.2, pp.280 – 289, (2006).

[5] Wataru Sunayama, Akihiro Iyama and Masahiko Yachida: HTML Text Segmentation for Web Page Summarization by Using a Key Sentences Extraction Method, Systems and Computers in Japan, John Wiley & Sons, Inc., Vol.37, No.7, pp.26 – 36, (2006).

[6] 砂山渡・矢田勝俊：説得プロセス分析の枠組みと債権回収会話ログへの適用，人工知能学会論文誌，Vol.22, No.2, pp. 239 – 247, (2007).

[7] 相良直樹・砂山渡・谷内田正彦：サブトピックを考慮した重要文抽出による報知的要約生成，電子情報通信学会論文誌，Vol.J90-D, No.2, pp.427 – 440, (2007).

[8] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸: 形態素解析システム『茶筌』, Version 2.2.9，使用説明書，(2002).