Vol.2010-CE-103 No.2
2010/3/6

データ結晶化を用いた対話ログの時系列解析

斎藤正孝[†1]　片上大輔[†1]　新田克己[†1]

本稿では，対話ログを時系列で区切り話題を出現単語のクラスタとして獲得し，その推移を視覚的に表示することで対話ログの概要把握の支援を行う時系列解析手法を提案する．時系列で区切ることで，全体では弱いが，局所的に見ると強くなるような単語の共起度を考慮し，従来手法より深い解析を行うことを目的とする．提案手法を用いた時系列解析システムを構築し，システムを利用しながら実際の対話ログを解析する．話題間の関係を抽出するデータ結晶化の技法を用い対話ログを解析した結果，話題の推移や関係を抽出することができた．

Time Series Analysis of Dialogue Logs with Data Crystallization

Masataka Saito,[†1] Daisuke Katagami[†1]
and Katsumi Nitta[†1]

In this paper, we introduce a time series analysis method of dialogue logs. Our method is to divide dialogue logs in chronological order and acquire the topics as clusters of appearance words by visualizing the process. Our goal is to brief the dialogue log and provide a deeper analysis than nonconventional method with regional co-occurrence by dividing dialogue logs. We develop the time series analysis system and analyze real dialogue logs with our system. After analyzing the dialogue logs by data crystallization which picks up the relation among topics, we succeed in picking up the transition and relation of dialogue logs.

1. はじめに

法学教育では，模擬調停，模擬交渉，模擬裁判等の交渉トレーニングにより，実践力を養い，対話や交渉の技術の向上に取り組んでいる．しかし，そのようなトレーニングは，教員は学生に対し対話の内容に関する助言を与える必要があり，同時に複数の学生がトレーニングを行うと部分的な対話を聞き助言を与えなければならない等，学生の対話の内容を全て把握して助言することができない．また，交渉トレーニングの記録を事後に評価する際も，人間が対話ログを直接読んで話題の推移や話題間の関係を把握するには，多大な労力と時間がかかってしまう．そこでコンピュータによる支援が必要とされてきた．

コンピュータによる対話ログの解析手法として，前野の「共起度に基づく単語クラスタリングを用いた話題抽出[1)]」がある．単語クラスタリングによる話題抽出では，単語の共起度を基にクラスタリングし，視覚化したグラフから話題を抽出することで，対話ログに出現する話題を抽出することができる．また，単語クラスタリング手法にデータ結晶化の手法を導入することで，話題間の関係を掴むことができることを示した．しかし，この手法は対話ログ全体で共起度を測定しているため，話題の推移が分からず，対話ログの支援としては不十分である．

前野は，自らのクラスタリング手法を改良し，時系列に沿ってクラスタリングする手法[2)]を提案した．しかし，この手法では，時系列を考慮しているものの，クラスタの中に複数の話題が含まれることがあり，正確に話題の推移を把握することは困難であった．

そこで，本稿では，対話ログを時系列で区切り話題を出現単語のクラスタとして獲得し，その推移を視覚的に表示することで対話ログの話題の推移や話題間の関係を把握する支援を行う時系列解析手法を提案する．時系列で対話ログを区切ることで，全体では弱いが，局所的に見ると強くなるような共起度を考慮し，従来手法より深い解析を行うことを目的とする．提案手法を用いた時系列解析システムを構築し，システムを利用しながら実際の対話ログを解析する．対話ログ全般が対象であるが，本稿では法学トレーニングに使用される，模擬調停・模擬交渉・模擬裁判といった論争に関する対話を扱う．先に挙げた対話では，話者は意図を隠しながら話を進めていくため，対話内容を把握するのは難しい．そのため，データ結晶化の技法を用い，対話ログを解析することで，本研究手法の有用性を示す．

2. 共起度に基づく単語クラスタリング

本章では，対話ログから話題を抽出し話題間の関係を掴む手法として，前野の「共起度に基づく単語クラスタリング[1)]」を説明する．

共起度に基づく単語クラスタリングでは，データ内に出現する単語の共起性を Jaccard 係数を用いて計測し，それに基づいて単語のクラスタリングを行い，それぞれの関係性を可視

†1 東京工業大学 大学院 総合理工学研究科

© 2010 Information Processing Society of Japan

Vol.2010-CE-103 No.2
2010/3/6

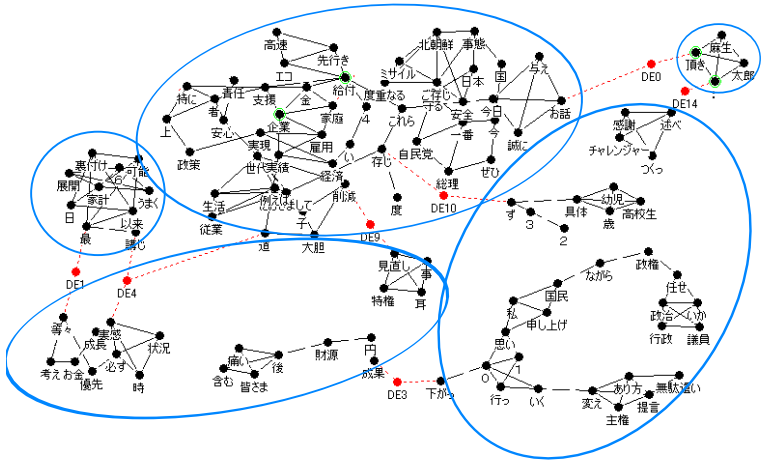

図 1 データ結晶化
Fig. 1 Data Crystallization

化したグラフを表示する．クラスタリングには k-medoids 法[3] を用いている．k-medoids 法ではクラスタ数を指定し，あらかじめ各クラスタにメドイドと呼ばれるクラスタの中心となる単語をランダムに配置し，それらの単語との関連性が高い単語同士を 1 つのクラスタとして得ることができる．

共起度を用いクラスタリングを行った結果は，単語の共起度による関係性をグラフに可視化する．グラフでは，単語が黒ノードとして表され，共起性が高いノード同士がリンクとして表される．表示されたグラフを観察して，対話の話題を抽出していく．

共起度に基づく単語クラスタリング手法では話題を抽出した後，話題間の関係を掴むために，データ結晶化[1)4)5)] の手法を用いている．データ結晶化とは，クラスタ間の関係を知るための手法である．データに含まれない単語をダミーノードとして発言に混入させ，クラスタ間に挿入することで，グラフ上でクラスタ間の関係を可視化する．この際に，発言記録中の各発言に対して以下の，ランキング関数[1)] を利用し，順位付けを行う．

$$I_{av}(b_i) = \frac{1}{|c|} \sum_{j=0}^{|c|-1} \min_{(e_k \in c_i) \wedge (e_k \in b_i)} Freq(e_k) \quad (1)$$

b_i はバスケット，c_i はクラスタ，$|c|$ はクラスタ数である．この値が大きい発言にダミーノードを割り当てて，グラフに追加する．

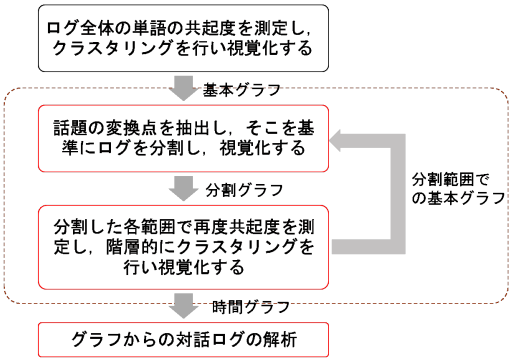

図 2 研究手法の概要
Fig. 2 Outline of research method

図 1 は，単語を 5 つのクラスタに分けた後，データ結晶化を用い，ダミーノードをグラフ上に可視化したものである．ダミーノード（DE0,DE1,DE3,DE4,DE9,DE10,DE14）は話題間の関係を示しているため，今話されている話題から別の話題へ移そうとしていたり，別の話題を引用しているなどの関係を表していると解釈できる．しかしながら，図 1 において，話題の推移をとらえるのは困難である．

3. 時系列クラスタリングの概要

本研究では，共起度による単語クラスタリング手法を改良し，階層的にクラスタリングを行う手法を提案する．本研究手法の流れを図 2 に示す．まず，従来と同様に単語クラスタリング手法を用い，対話ログ全体の**基本グラフ**を表示する．そして，話題の転換点でログを分割し，それぞれ分割した範囲で**分割グラフ**を表示する．分割範囲で再度共起度を測定し，クラスタリングを行うことを繰返すことでログを分割していく．最後に分割グラフを結合することで**時間グラフ**を表示し，グラフを見ることでログの話題の推移や話題間の関係を見ていく．階層的に分割していき，時には分割した範囲を結合しながら，解析を進めていく．

本研究では，時系列クラスタリングを行う上でグラフを 3 つ用意した．以下に，それらを示す．

© 2010 Information Processing Society of Japan

基本グラフ

基本グラフは，共起度による単語クラスタリング手法と同じアルゴリズムで表示されるグラフである．基本グラフは解析範囲全体の単語の共起性を可視化する．対話ログを分割し，その分割した範囲の基本グラフを表示させることで，局所的な共起性を考慮し，全体の基本グラフより，まとまりのある話題が抽出できる．

分割グラフ

分割グラフは，対話ログを分割し，それぞれの分割範囲に応じたノードの出現度を濃淡として表現したグラフである．基本グラフと形は全く同じだが，ノードの濃淡のみが異なるグラフとなる．黒ノードとダミーノードの両方の濃淡を変化させるが，双方の濃淡を示すアルゴリズムは異なる．分割グラフを表示するには，まず分割範囲を指定する．黒ノードの濃淡は，解析範囲内に出現するノード数に対する分割範囲内に出現するノード数の割合に依存する．ある閾値より高い割合を示す黒ノードは基本グラフと同様に表示し，低い黒ノードを薄く表示することで濃淡を表現する．

ダミーノードは，分割範囲に含まれるバスケットに挿入されたダミーノードを基本グラフと同様に濃く表示し，それ以外のダミーノードを薄く表示する．

分割グラフで，濃く現れたクラスタは，その分割範囲で出現頻度が多いため，クラスタに対応した話題がより多く話されていたことになる．つまり，分割グラフでクラスタの濃淡を見ることで，分割範囲で話されていた話題を把握することができ，時系列順に分割グラフを見ることで，話題の推移を把握することができる．

時間グラフ

時間グラフは，時間的に連続する分割グラフを縦に並べ，その間にダミーノードを挿入したものである．ダミーノードは，連続する分割範囲の分割位置付近に挿入されたダミーノードを表示させる．

連続する分割範囲で重なる部分は，話題の転換点であると考えられるため，この部分に挿入されたダミーノードには通常のダミーノードよりも意味があると思われる．これらのダミーノードを区別するために，時間グラフにおいて，分割グラフの中に表示されているダミーノードを内部ダミーノード，分割グラフの間に表示されているダミーノードを外部ダミーノードと呼ぶ．

4. 時系列解析システム

本章では，開発した時系列解析システムの概要を示す．なお，システムは瀬々[6] のシステムを拡張することで開発した．

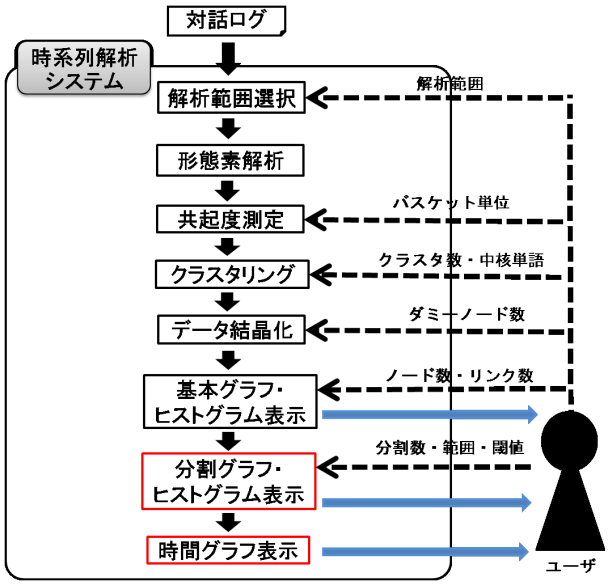

図 3 システムの構成図
Fig. 3 System configuration diagram

本システムの構成を図 3 に示す．黒の実線はシステムの処理の流れで，青の実線はシステムのユーザに対する視覚化，点線はユーザのパラメータ操作を表している．文字が赤くなっている部分が拡張した部分である．

以下，システムの処理の流れに沿って説明をする．

(1) 解析範囲選択

対話ログを読み込み，ユーザが指定した解析範囲で対話ログを区切る対話ログ全体ではなく，一部を切り出して局所的に解析したい場合にも解析範囲を選択する．

ユーザは解析範囲の開始部分と終了部分の id 番号を選択することで，選択した部分のみのログを解析することができる．選択範囲以外のログは全く使わない．

(2) 形態素解析

形態素解析では，入力文書を言語学的に意味をもつ最小単位である形態素に分解する．各形態素の品詞を決定すると共に，活用などの語形変化をしている形態素に対しては原形を割

© 2010 Information Processing Society of Japan

り当てることができる．原形の形態素をノードとしてグラフに出力する．ここでは，奈良先端科学技術大学院大学で開発された形態素解析ツール ChaSen[7] を利用した．また，ストップワードとして，助詞・助動詞・代名詞・連体詞・接頭語・記号を除去した．

(3) 共起度測定

形態素解析によって生成された単語を用い，ユーザが指定したバスケットを単位として単語の共起度を測定する．共起度を測定するバスケット単位は，発言単位か文単位の 2 つであり，ユーザがどちらか好きな方を選択する．

(4) クラスタリング

ユーザが指定したクラスタ数とクラスタの中核となる単語を基に，瀬々のクラスタリングの拡張手法[6] を用い，話題とクラスタが対応するような単語クラスタリングを行う．

ユーザはパラメータとして，クラスタ数とクラスタ中核単語を指定する．クラスタ中核単語は各クラスタごとに，0 個以上の単語を指定することができる．システムは，クラスタ中核単語を選ぶ時の基準として，解析範囲に出現する単語を頻度により降順に並べたリストをユーザに示す．選択されたクラスタ中核単語が，k-medoids 法の初期単語として与えられる．

(5) データ結晶化

ユーザが指定したダミーノード数分，クラスタ間にダミーノードを挿入する．共起度測定で指定したバスケットに対してランキング関数[1] を使用し，バスケットに順位付けを行った後，ユーザが指定したダミーノード数分，上位のバスケットにダミーノードを挿入する．次に，異なるクラスタ間で共起度の高い単語を選び，それらの単語とダミーノードをリンクする．

(6) 基本グラフ表示

基本グラフを表示する．表示する黒ノードと黒リンクの数はユーザの指示による．黒ノードは出現頻度，黒リンクは共起度の上位からユーザが指定した分選択する．グラフ中の赤ノードはダミーノードで，ダミーノードに付随するリンクは赤の点線で表す．ダミーノードに付けられた番号は，ダミーノードを挿入したバスケット番号（発言 id）である．

また，解析範囲内で，各クラスタに含まれる単語の頻度の移り変わりを示すヒストグラムを表示する．ヒストグラムの例を図 5 に示す．ヒストグラムの横軸は解析範囲に含まれる発言 id で，縦軸は各クラスタに属する単語の頻度の総和を表している．横軸の間隔は初期値として 5 が与えられているが，ユーザは好きな間隔を指定することができる．

クラスタは話題に対応しているため，ヒストグラムでクラスタ内のノード数の時間的な推移を見て取ることで，ユーザは話題の推移を推測することができる．クラスタ内のノード数が増減している部分で，対応する話題の発言数も増減している．特に，ヒストグラムの線の交差点では，一方のクラスタが増え，もう一方のクラスタが減少しているので，そこで話題が切り替わっていると考えることができる．ユーザは，線の交差点を参考にしながら，グラフを時間分割する発言 id を決定する．

期待するグラフやヒストグラムが得られなかった場合は，(1)〜(5) に戻りパラメータを再設定し，再び基本グラフを表示する．

(7) 分割グラフ表示

ヒストグラムを参考にし，ユーザは時間分割する数と各分割範囲の開始 ID と終了 ID を指定する．そして，指定された分割数・分割範囲を基に，それぞれのクラスタに含まれる単語の出現頻度を計算し，出現頻度に応じた濃淡を示すグラフを表示する．分割グラフを表示する際，黒ノードの濃淡を表す閾値は，0〜100 の範囲での百分率の割合をユーザが指定することができる．

分割グラフを用いることで，グラフの濃淡の推移が見て取れる．グラフのノードが濃い部分が，分割範囲内で出現率の高いノードであるため，グラフの各クラスタの濃淡に着目することで，分割範囲での話題や，話題の推移を把握することができる．クラスタ間の濃淡がきれいに分かれない場合は，ヒストグラムを用い，分割範囲や閾値を微調整する．

(8) 時間グラフ表示

連続する分割グラフを縦に表示し，間にダミーノードを挿入することで時間グラフを表示する．外部ダミーノードは，分割グラフの間，つまり話題の転換点に挿入されているダミーノードなので，不利な話題から有利な話題に移そうとする意図等，話題の転換に関係した性質を持つダミーノードであると考えることができる．一方，内部ダミーノードは話題の中に出現してくるダミーノードなので，別の話題の引用等，話題は移行しないが他の話題が出現するような性質を持つダミーノードであると考えることができる．外部ダミーノードと内部ダミーノードに着目することで，ユーザは更なる解析のきっかけを得ることができる．

5. 解　析　例

本章では，開発したシステムで解析を行うことで，解析例を示す．

5.1 解析に用いた調停事例

対話ログとして，車のマフラーのネットオークショントラブルの模擬調停の記録を使用した．この調停では，出品者がノークレーム・ノーリターンとして車のマフラーを出品し，落札者がそのマフラーを落札した．落札者はステンレス製のマフラーを期待して落札した

© 2010 Information Processing Society of Japan

のだが，届いたマフラーは欲しかったステンレスより品質の劣るアルスター製のものであった．出品者は，材質についてはメーカーのホームページの URL を載せたのみで，そのホームページに載っているマフラーは全てステンレス製のものであった．また，通常は刻印されているはずの製造番号が，このマフラーには無かった．そこで落札者は返品するので返金して欲しいと要求し，商品が届いてから 2ヶ月後にその旨を伝えるメールを送った．お互いに問題があるため，それぞれ相手の評価を「悪い」にしている．

この調停では，様々な論点が考えられるが，対話ログで実際に現れた論点は以下の通りである．

- 材質…ホームページ上にはステンレス製のマフラーしか載っていなかったが，落札したマフラーはステンレスより劣るアルスター製のものだった
- 刻印…通常なら製造番号の刻印が存在するが，落札したマフラーには刻印がなかった
- ノークレーム・ノーリターン…出品者はノークレーム・ノーリターンで出品したが，落札者はそれに応じようとしない
- 返金…調停案の 1 つ．落札したマフラーを返品する代わりに返金を要求した
- 交換…調停案の 1 つ．落札したマフラーを，出品者が所持する別のステンレス製のマフラーと交換する

5.2 解析手法

対話ログの全範囲を解析した基本グラフ（図 4）を表示する．5 つのクラスタに属する単語から，各クラスタは，材質・ノークレームノーリターン・刻印・交換・返金の話題に対応していると読み取れる．次に，ヒストグラム（図 5）を基に分割範囲を指定した．このヒストグラムで線は様々な箇所で交差するが，交差点が密集している部分で特に話題の推移があると考えられるため，特に密集している部分で 4 つの区間に分割した．ここでの分割範囲はそれぞれ ID0〜ID21，ID15〜ID42，ID37〜ID70，ID65〜ID91 である．それぞれの分割範囲で分割グラフを表示したものが図 6 である．

この分割グラフでは，まだ話題のおおまかな推移しか分からないので，更に分割範囲を解析していく．ここでは，特にクラスタの濃淡がまばらであった分割範囲 2 を解析していく．

分割範囲 2 を解析範囲として，先ほどと同様に基本グラフ・ヒストグラムを表示し，分割グラフを表示させたものが図 7 である．このグラフを見ると，青で囲んだクラスタが特に濃く現れているため，各分割範囲で対応した話題が多く話されていることが分かる．この範囲では，材質→交換→刻印と話題が移行していったと読み取れる．

各範囲に同様の解析を行い，最終的に分割した範囲の分割グラフを時系列ごとに読み取る

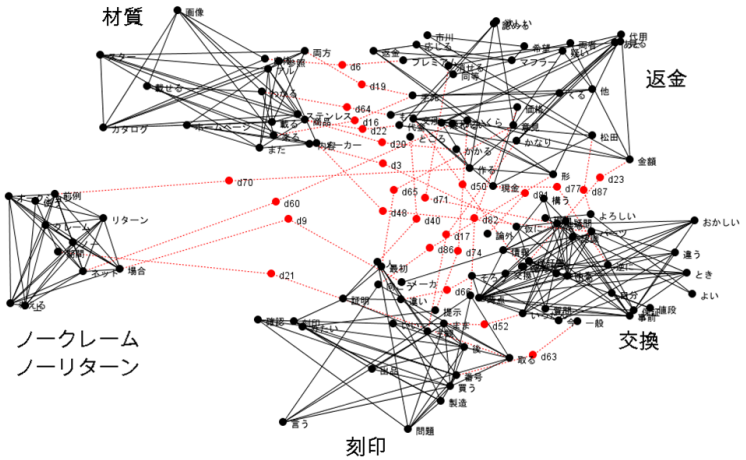

図 4 基本グラフ
Fig. 4 Base Graph

ことで，対話ログ全体の話題の推移を見ていく．この解析では，最終的に 11 個の範囲に分割できた．

このように話題の推移が読み取れたら，時間グラフを表示する．分割範囲 2 を解析した時の時間グラフを図 8 に示す．時間グラフの外部ダミーノードと内部ダミーノードの解釈を比較することで，それぞれの特徴を観察する．

5.3 話題の推移

分割グラフのクラスタの濃淡を読み取ることで，各分割範囲に対応している話題を読み取り，帯グラフに表示したものが図 9 である．上の帯グラフが全範囲の解析から読み取れたもので，4 つの分割グラフから話題を読み取った．中央の帯グラフが階層的な時系列解析から読み取れたもので，11 個の分割グラフから話題を読み取った．なお，分割範囲 1 を更に 3 つに分割した範囲をそれぞれ，1-1,1-2,1-3 と表現している．下の帯グラフは対話ログでの実際の話題の推移である．これらのグラフを比較することで，話題の推移の考察とする．

まず，上の 2 つの帯グラフを比較し，時系列を考慮していない解析と考慮している解析の違いを見る．全範囲の解析から読み取れた話題の推移では，全体の共起度の測定のみでグラフを表示しているため，単語が広い範囲で共起性を持ってしまい，うまく話題が読み取れていないことが分かる．つまり，分割せずに対話ログ全体のグラフを表示しただけでは，

© 2010 Information Processing Society of Japan

Vol.2010-CE-103 No.2
2010/3/6

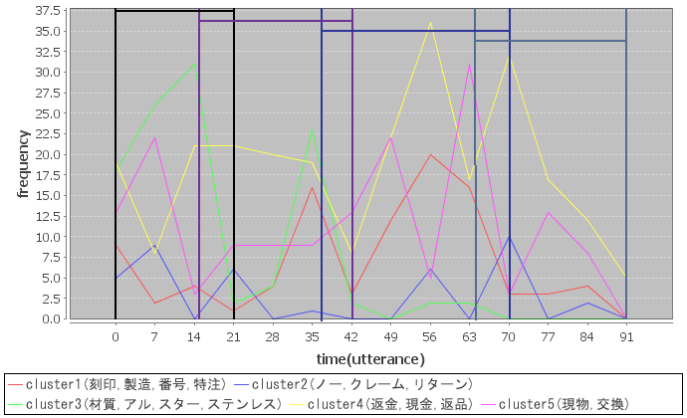

図 5 ヒストグラム
Fig. 5 Histgram

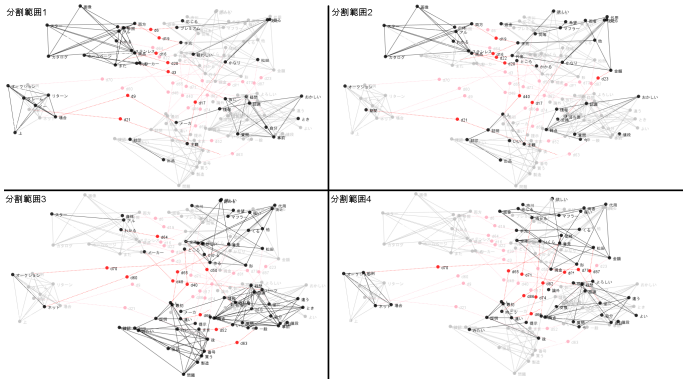

図 6 対話ログ全体を解析した時の分割グラフ
Fig. 6 Divide Graph by analyzing entire range

話題の推移を読み取ることは困難である．しかし，最終的に得られた分割グラフでは，局所的な共起度を考慮し，各話題に対応した単語が同じクラスタに分類されたため，うまく話題を読み取れた．なお，2 つ話題が読み取れている部分があるが，これは更に分割しても，きれいに話題ごとに分かれなかったため，この部分は話題が 2 つであると判断している．

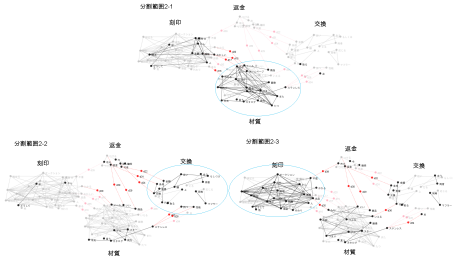

図 7 分割範囲 2 を解析した時の分割グラフ
Fig. 7 Divide Graph by analyzing divide range 2

図 8 分割範囲 2 を解析した時の時間グラフ
Fig. 8 Time graph by analyzing divide range 2

次に，下の 2 つの帯グラフを比較し，階層的な時系列解析から読み取れた話題の推移と，実際の話題の推移を見比べると，ほぼ話題の推移が一致していることが分かる．本研究手法である，対話ログを分割し階層的にクラスタリングして視覚化する手法を用いることで，話題の推移が読み取れた．

5.4 ダミーノード

本節では，外部ダミーノードと内部ダミーノードを比較することで，それぞれの特徴を考察していく．分割範囲 2 に出現した外部ダミーノードと内部ダミーノードを見ていく．

5.4.1 外部ダミーノード

分割範囲 2 の時間グラフを見ると，ダミーノードが 4 つ出現したが，ここでは d31 に注目する（図 10）．

d31 は，返金と交換のクラスタに繋がれているため，返金の話題と交換の話題に関係があるダミーノードである．ここで，実際に d31 が混入されたバスケットである 31 番目の発言を見てみると，30 番目の発言で調停者が返金に関する話を振っているが，31 番目の発言で，出品者は交換なら応じられるが返金には応じられないと言っている．出品者にとって，返金の話題は都合が悪く，交換の話題の方が都合が良いため，都合の悪い話題から話題をそらそうとしている意図が伺える．

© 2010 Information Processing Society of Japan

Vol.2010-CE-103 No.2
2010/3/6

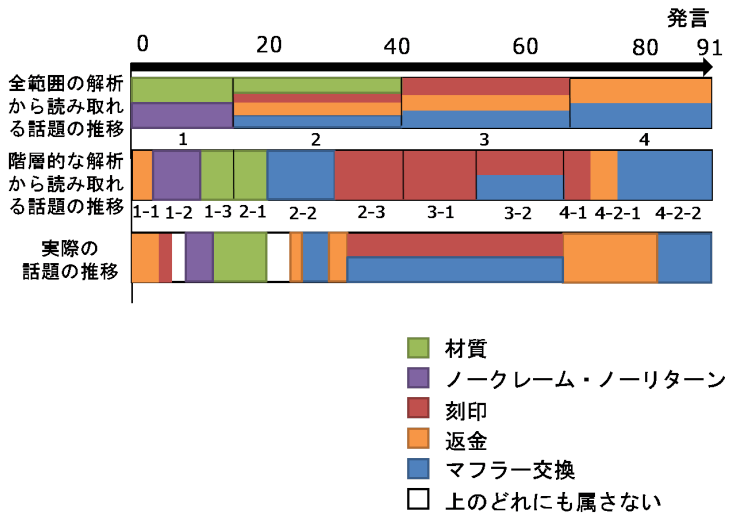

図 9 分割グラフから読み取れる話題と実際の話題の推移
Fig. 9 Transition of topics from Devide Graph and a dialogue log

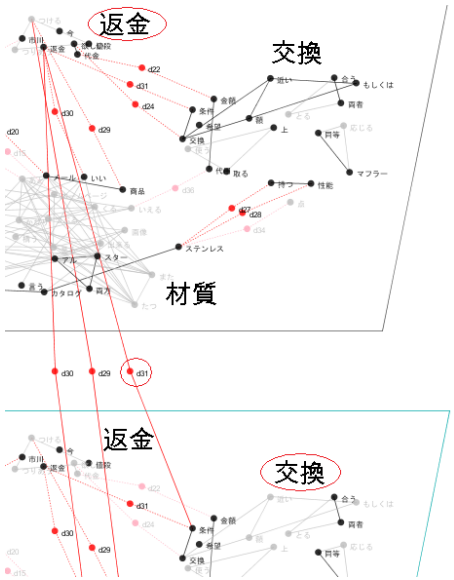

図 10 外部ダミーノード d31
Fig. 10 External dummy node d31

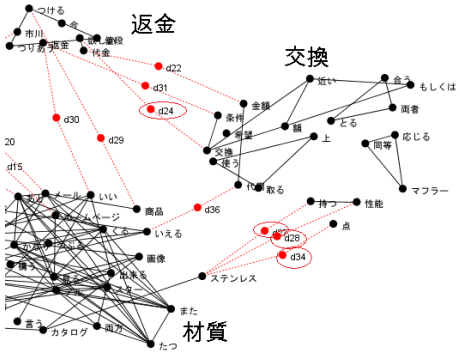

図 11 内部ダミーノード d24,d27,d28,d34
Fig. 11 Internal dummy node d24,d27,d28,d34

5.4.2 内部ダミーノード

同じ分割範囲 2 で出現した内部ダミーノードを見てみる．内部ダミーノードは数多く出現したが，ここでは，d24,d27,d28,d34 に注目する．

d24 は，返金と交換のクラスタに繋がれているため，返金の話題と交換の話題に関係があると考えられる．実際に発言 24 を見てみると，調停者が返金と交換を引き合いに出している．

また，d27,d28,d34 は，材質と交換のクラスタに繋がれている．実際にこれらの発言を見てみると，交換の話題で「ステンレス製」という材質の話題を引用していることが分かる．

以上より，外部ダミーノードは，不利な話題から有利な話題へ移行しようとする話題の推移に関する関係が，内部ダミーノードは，話題の引き合いや引用といった関係が読み取れた．外部ダミーノードは話題の転換点に出現するダミーノードなので，内部ダミーノードに比べて話題の推移を促す発言が観察できた．

6. おわりに

本稿では共起度による単語クラスタリング手法の改良として，対話ログを時系列で区切り話題を出現単語のクラスタとして獲得し，その推移を視覚的に表示することで対話ログの概要把握の支援を行う時系列解析手法を提案した．時系列で区切る際に，ヒストグラムの交差点を基準とすることで，話題の転換点で分割することができ，分割範囲で再度共起度を測定しクラスタリングすることで，分割範囲の局所的な共起度を考慮できていることが分かった．

また，階層的なクラスタリングを行った結果，分割グラフのクラスタの濃淡から話題を読み取ることで，話題の推移を見て取ることができた．これは，時系列を使用しない単語クラスタリングや他の時系列クラスタリングでは読み取れない，対話にとって重要なことである．

そして，外部ダミーノードと内部ダミーノードに分けて，それぞれの傾向を見ることで，

© 2010 Information Processing Society of Japan

Vol.2010-CE-103 No.2
2010/3/6

外部ダミーノードからは話題の推移に関する発言が，内部ダミーノードからは話題の引き合いや引用をしている発言が観察できた．

謝　　辞

研究のベースとなる「共起度に基づく単語クラスタリング手法」を教示してくださった東京大学工学研究科の大澤幸生教授，ソーシャルデザイングループの前野義晴氏にお礼申し上げます．

参 考 文 献

1) 前野義晴，コミュニケーションから探る組織の見えない黒幕，人工知能学会論文誌 Vol.22 No.4 pp.389-396 (2007)
2) Y.Maeno, Reflection of the agreement quality in mediation JURISIN2009 pp.73-82 (2009)
3) T.Hastie,R Tibshirani, and J. Friedman: The elements of statistical learning - Data mining,inference,and prediction Springer-Verllag (2001)
4) 大澤幸生，チャンス発見のデータ分析，東京電機大学出版局 (2006)
5) Y.Ohsawa, Data crystallization:chance discovery extended for dealing with unobservable events, New Mathematics and Natural Computation vol.1 pp.373-392 (2005)
6) 瀬々佳奈，KeyGraph とデータ結晶化を用いた発言ログからのシナリオ抽出支援システム, 電子情報通信学会 (2007)
7) http://chasen.naist.jp/hiki/ChaSen/

© 2010 Information Processing Society of Japan