人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第9回)
SIG-AM-09-02

# 協調的マルチビューに基づくインタラクティブ文書クラスタリングシステムの提案

利根川 拓馬* 高間 康史

Takuma Tonegawa, Yasufumi Takama

首都大学東京大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

**Abstract:** 本稿では，協調的マルチビューに基づくインタラクティブ文書クラスタリングシステムを提案する．提案システムでは，ユーザフィードバックをクラスタリング結果に反映するために，単語の重み調整に基づく手法を採用し，クラスタや文書，単語と言った異なるレベルの情報を効率的に提示するために協調的マルチビューを採用する．TETDM（Total Environment for Text Data Mining）を用いてプロトタイプシステムを実装し，評価実験を行った結果について示す．

## 1. はじめに

本稿では，協調的マルチビューに基づくインタラクティブ文書クラスタリングシステムを提案する．学術論文や最新のニュース記事などの文書データは，様々な知識を得るための重要なリソースである[1]．近年，それらの文書データに対して，話題検出・追跡や文書間の関係性の発見などの探索的データ分析を行う必要性が高まっている[9]．そのため，ユーザの探索データ分析を支援し，負担を軽減することを目的としたインタラクティブテキストマイニングシステムの研究が進められている．

探索的データ分析の代表的手法の一つにインタラクティブクラスタリング[2]がある．教師なし学習である通常のクラスタリングとは異なり，インタラクティブクラスタリングではユーザがオブジェクトをグループ化する際にいくつかの制約を与えるため，半教師あり学習と呼ばれる．これにより，ユーザの視点を反映させたクラスタリングを行うことができ，効率よくデータの分析を行うことが可能となる．しかし，インタラクティブクラスタリングシステムを開発する際には,「複数オブジェクトの情報をどのように表示するか」,「異種オブジェクト間の関係性をどのように表示するか」,「制約付きクラスタリングをどのように導入するか」といった問題が挙げられる．

* 連絡先:首都大学東京大学院システムデザイン研究科
〒191-0065 東京都日野市旭が丘 6-6
E-mail:ytakama@sd.tmu.ac.jp

これらの問題に対し有効なアプローチとして，本稿では協調的マルチビュー（Coordinated Multiple Views，CMV）のコンセプトに着目する．提案システムでは，ユーザに提示すべき情報をクラスタレベル，文書レベル，単語集合レベル，単語レベルの 4 レベルに分け，それぞれを別のビューに表示することで，複数種オブジェクトの情報を適切なビューに表示することができる．

また，ビュー間の協調を可能にすることで，異種オブジェクト間の関係性を把握可能とする．制約付きクラスタリング手法としては，類似度計算における単語の重みをユーザが制約として与える手法を採用する．さらに，各文書に対してユーザによるラベル付けを可能とすることで，ユーザの視点とクラスタリング結果の比較を支援する．

提案システムの開発には，テキストデータマイニングのための統合開発環境（TETDM）[3]を採用する．上述のクラスタレベル，文書レベル，単語集合レベル，単語レベルそれぞれに対応したパネル，およびパネル間の連動を実装する．TETDM を用いて実装した提案システムを用いて，言語処理学会[1]年次大会の論文データを対象とした評価実験を行った結果を示す．

## 2. 関連研究

### 2.1 テキストデータマイニング

膨大な文書データから有用な情報を発見したり，

[1] http://www.anlp.jp/

文書データ間の関係性を把握したりするといった，ユーザの様々な要求に十分に対応できる情報アクセス手段の重要性が指摘されており[4]，情報抽出，文書検索，文書分類や文書クラスタリングなどのテキストデータマイニングの技術が研究されている．

本研究で利用するテキストデータマイニング統合環境 TETDM は，柔軟な方法で様々なテキストマイニング技術を組み合わせることを目的として開発されている[3]．データ分析ツールを開発し，統合環境内にモジュールとして組み込むことが可能である．モジュールには文書処理を実行する処理モジュールと，処理モジュールの出力を表示する可視化モジュールの 2 種類がある．TETDM はインタフェース画面に設置されたパネルに対し，処理モジュールと可視化モジュールを 1 対 1 で組み合わせることでツールを構成する．また，TETDM は連動処理部によってモジュール間の連携を制御・実施することが可能であり，異なるパネル間の協調の実装を効率的に行うことができる．

## 2.2 協調的マルチビュー

複数のビューから構成される可視化システムを設計するために，協調的マルチビューのコンセプトが提案されている[5]．複数のビューによって提示された情報が，協調によって相互作用することで，ユーザはデータを効率良く理解することが可能となる．

代表的なマルチビューのタイプの一つに，一方のビューでデータの全体もしくは非常に大きな部分（overview）を表示し，別のビューでデータのより詳細部分（detail view）を表示する，Overview + Detail views がある．Zhang ら[6]は，このタイプのマルチビューを用いてインターネットログの異常を検出するネットワーク管理システムを提案している．

一般的な協調に Brushing と Navigational Slaving がある. Brushing は, あるビューで要素を選択すると，リンクされた他のビューにおける同一の（もしくは関連のある）要素が同時にハイライトされる．Navigational Slaving はユーザがあるビューでスクロールなどのナビゲーション動作を行うと，リンクされた他のビューに自動的に反映される事を指す．Weaver[7]は，Brushing や Navigational Slaving による協調機能を持ったマルチビューをユーザがインタラクティブに構築可能なシステムを提案している．

# 3. 提案システムの概要

## 3.1 提案システムのコンセプト

インタラクティブクラスタリングシステムを開発する際,「複数オブジェクトの情報をどのように表示するか」,「異種オブジェクト間の関係性をどのように表示するか」,「制約付きクラスタリングをどのように導入するか」といった点を検討する必要がある．前者の 2 点に関して，本稿では協調的マルチビューのコンセプトを採用する．提案システムを設計するにあたり，以下の 4 点を考慮する．

・文書情報を 4 つのレベルに分けて並列表示：各レベルの情報を適切なビューに表示
・ビュー間の協調：異種オブジェクト間の関係性を提示
・任意の単語の重み変更，再クラスタリングを反復的に実行：ユーザの視点を反映させた制約付きクラスタリングの導入
・任意の文書へのラベル付与：ユーザの視点とクラスタリング結果の比較を支援

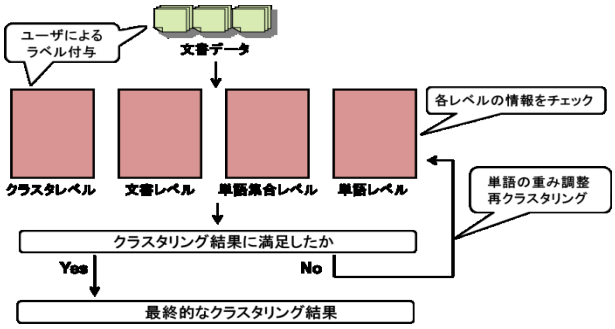

図 1：提案システムのフローチャート

図 1 に提案システムのフローチャートを示す．ユーザは関心を持った文書にラベルを付与し，クラスタリング結果を調べる際に利用する．

提案システムはクラスタリング結果についての情報を 4 つのレベル（クラスタレベル，文書レベル，単語集合レベル，単語レベル）に分け，各レベルの情報を異なるビューに並列表示する．これらのビューを組み合わせることで，ユーザは効率よくクラスタリング結果を確認することができる．表 1 に各レベルに提示する情報及び可能な操作を示す．

文書クラスタリングには k-means アルゴリズムを採用する．文書$d_i = (w_{i1}, \cdots, w_{in})$における単語$t_j$の重み$w_{ij}$は tf-idf 値を用いる．

提案システムは単語の重みによってユーザフィードバックを与えるインタラクティブクラスタリングを採用する[8]．ユーザが単語の重みを調整したい場合, 任意の単語の重みに$3^k$を掛けることによって単語の重みを調整し，システムにフィードバックを与える．$k$の値は単語レベルに対応したパネルでインタラクティブに変更することができる．全単語の$k$の値の初期値は 0 となっており，1 ずつ増減可能である．文書間類似度は式(2)によって計算する．式(2)

はコサイン類似度を元にしており，調整された単語の重みを文書間の類似度に反映させている．

$$sim(d_1,d_2)=\frac{\sum_{j=1}^{n}3^{2kj}w_{1j}w_{2j}}{\sqrt{\sum_{j=1}^{n}(3^{kj}w_{1j})^2}\sqrt{\sum_{j=1}^{n}(3^{kj}w_{2j})^2}} \quad (2)$$

重み調整と再クラスタリングを反復的に行うことによって，ユーザは最終的に満足のいくクラスタリング結果を得ることができる．また同時に，クラスタリング結果に影響を与える単語を理解することや，各クラスタリング結果から新たな知識を得ることが可能となる．

表 1：各レベルに提示する情報及び可能な操作

| レベル | 提示する情報 | 可能な操作 |
|---|---|---|
| クラスタレベル | ・クラスタリング結果<br>・各文書のタイトル<br>・重みが調整された単語とその重み | ・クリックによる文書番号の指定<br>・クラスタ数の変更<br>・文書へのラベル付与 |
| 文書レベル(指定文書に関する情報) | ・指定文書本文<br>・指定文書の重要文<br>・指定文書の重要単語<br>・指定文書の情報(文数，単語数) | |
| 単語集合レベル | ・指定文書の単語一覧とその出現頻度<br>・指定文書の単語間のコサイン類似度 | ・クリックによる単語の指定 |
| 単語レベル(指定単語に関する情報) | ・指定単語を含む文書一覧<br>・指定単語の意味<br>・指定単語とコサイン類似度の高い単語 | ・クリックによる文書番号の指定<br>・指定単語の重み調整<br>・再クラスタリング |

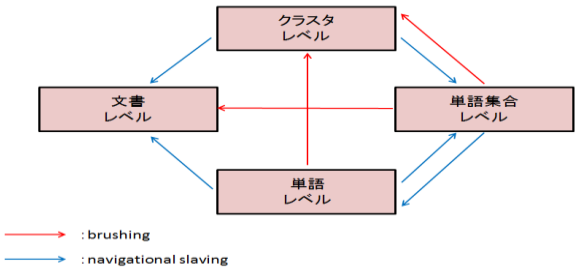


図 2：各レベル間の協調

図 2 に提案システムの各レベル間での協調を示す．図 2 において，赤い矢印は Brushing に対応し，青い矢印は Navigational Slaving に対応している．ユーザがクラスタレベルに対応したビューで文書を指定した場合，その情報が文書レベルと単語集合レベルに対応したビューに表示される．また，ユーザは単語レベルに対応したビューでも文書を指定することができ，指定された文書に関する情報が文書レベルと単語集合レベルに対応したビューに表示される．さらに，単語レベルに対応したビューで文書を指定した場合，その文書を含むクラスタがクラスタレベルでハイライトされる．

ユーザが単語集合レベルに対応したビューで単語を指定すると，その情報が単語レベルに対応したビューに表示される．さらに，指定単語が文書レベルに対応したビューに表示されている本文中に出現している場合，その部分がハイライトされる．同時に，クラスタレベルに対応したビューにおいて指定単語を含む文書番号もハイライトされる．

## 3.2 提案システムのプロトタイプ

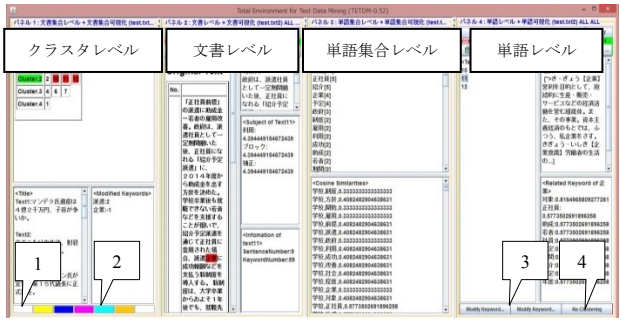


図 3：提案システムのインタフェース

図3に提案システムのインタフェースを示す．提案システムは4つのパネルから構成されており，左端のパネルから順に，クラスタレベル，文書レベル，単語集合レベル，単語レベルにそれぞれ対応している．また，クラスタ数変更テキストフィールド（図中1）でクラスタ数の変更，ラベル付与テキストフィールド（図中2）で文書へのラベル付与，単語の重み調整ボタン（図中3）で単語の重み調整，再クラスタリングボタン(図中4)で再クラスタリングが可能である．

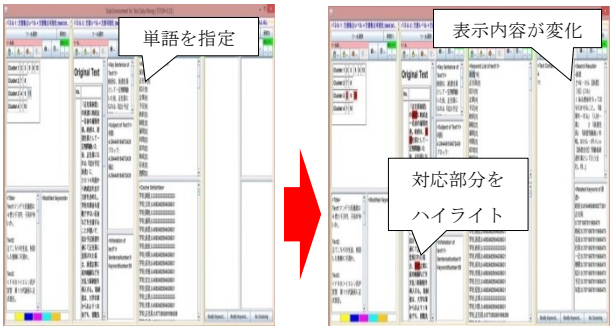


図 4：協調の例

図 4 に協調の例を示す．ユーザが単語集合レベルに対応したパネルにおいて，単語をクリックすると，指定された単語の情報が単語レベルに対応したパネルに表示される．また，クラスタレベルに対応したパネルのクラスタリング結果で，指定された単語を含む文書番号が赤色でハイライトされる．同時に，指定単語が文書レベルに対応したパネルに表示されている文書の本文中に出現している場合，その部分が赤色でハイライトされる．

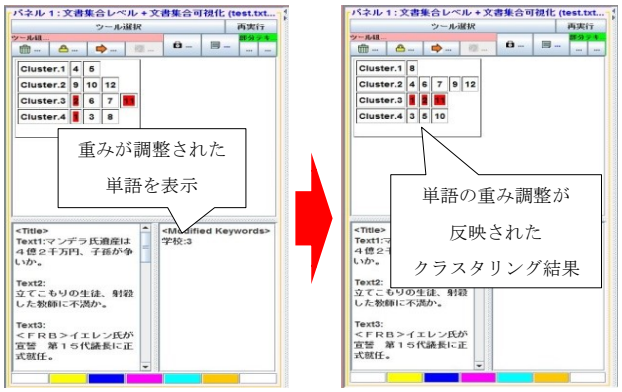


図 5：単語の重み調整と再クラスタリングの例

図 5 に単語の重み調整と再クラスタリングの例を示す．ユーザは各レベルの情報を元に，重みを調整したい単語を決定し，単語レベルに対応したパネルにある単語の重み調整ボタンをクリックする．ユーザが単語の重み調整ボタンをクリックするごとに，前述の単語の重み調整係数である $3^k$ の $k$ の値が 1 ずつ増減し，指定した単語の重み調整が実行される．同時に，重みが調整された単語とその重みがクラスタレベルに対応したパネルに表示される．ユーザが単語レベルに対応したパネルの再クラスタリングボタンをクリックすると，クラスタリングが再実行され，クラスタレベルに対応したパネルのクラスタリング結果が更新される．

単語の重み調整と再クラスタリングを反復して行うことにより，最終的にユーザの望むクラスタリング結果を得ることができ，ユーザの視点を反映させた制約付きインタラクティブクラスタリングシステムを実現している．

また，クラスタレベルに対応したパネルのラベル付与テキストフィールドにおいて，利用したい色のテキストフィールドを選び，ラベルを付与したい文書番号を入力すると，クラスタレベルに対応したパネルに表示されるクラスタリング結果において，入力した文書番号の色が指定した色に変更される．

# 4. 評価実験

## 4.1 実験概要

文書クラスタリングにおける提案システムの有用性や，ユーザにとって有用な情報や協調，また，提示する情報の違いがユーザの分析作業や実験結果に与える影響を分析するために，工学系の大学生および大学院生の男女16名に協力を依頼し評価実験を行った．実験では，言語処理学会年次大会発表論文集の，2002年と2003年におけるポスター発表の予稿データを利用し，「一方の年次に特有の話題」および「両方の年次に共通の話題」を発見するとともに，発見した話題に関するクラスタを生成するタスクを行ってもらった．また，提示する情報の違いが分析作業に与える影響について調査するために，単語集合レベルに対応したパネルの有無によって，実験協力者を8人ずつの2グループに分けて実験を行った．実験終了後には，発見した話題と提案システムの提示情報・機能などの有用度に関する5段階評価のアンケートに回答してもらった．また，視線追跡装置 Tobii X120[2]を用いて記録した，作業中の実験協力者の視線データの分析も行う．

## 4.2 実験結果

表 2：発見された話題

| | | 単語集合レベルあり | 単語集合レベルなし |
|---|---|---|---|
| クラスタが形成された話題 | 共通話題 | ・特許(3)<br>・手話(16)<br>・翻訳(18)(21)(18)<br>・日本語の文節解析(21)<br>・言語の分類(16)<br>・言語の分析(26)<br>・話し言葉(6) | ・検索(21)<br>・翻訳(15)(37)<br>・電子テキスト(29)<br>・学習支援を目的とした研究(19)<br>・ユーザを想定した研究(11)<br>・手話(12)<br>・テキスト(17)<br>・インターネット検索(20)<br>・2言語間の変換(14)<br>・コーパス(78) |
| | 特有話題 | ・機械翻訳(35)<br>・web サイトを利用した研究(6)<br>・音韻(2)<br>・外国語の音声翻訳(3)<br>・品詞(4)<br>・形態素解析(7)<br>・話し言葉(6)<br>・換言(4) | ・SVM(3)<br>・Perl を使用した研究(4)<br>・テストコレクションの利用(2)<br>・通訳(2)<br>・自動的(8)<br>・発話(4) |
| クラスタが形成されなかった話題 | 共通話題 | ・検索<br>・要約支援<br>・異なる言語の処理<br>・翻訳<br>・対話<br>・データベース<br>・談話<br>・自然言語処理<br>・音声(対話)<br>・文法 | ・音声<br>・機械翻訳<br>・談話<br>・言語解析<br>・運転<br>・音声翻訳<br>・音声認識 |
| | 特有話題 | ・形態素解析 | ・中国語<br>・音声解析・利用<br>・言葉 |

表 2 に発見された話題を示す．括弧内の数字はその話題に対応したクラスタのサイズ（論文数）を示している．実験協力者あたりの平均発見話題数を表

[2] http://www.tobii.com/

3 に示す．表 3 より，発見した話題の数は共通話題の方が多い．これはデータセットの年次が近いためと考える．また，単語集合レベルがある場合の方が，発見した話題の数が多くなっており，単語集合レベルのパネルによって，実験協力者が単語に注目しやすく，効率良く話題を発見可能であったと考える．

表 2 より，特有話題のほとんどはクラスタ形成されており，そのクラスタサイズは小さい．これより，関連文書数が少ないほど，重みの調整によるフィードバックでクラスタにまとめやすいと考える．逆に，共通話題はクラスタが形成されなかった話題が多く，クラスタが形成されている話題についてもクラスタサイズが大きくなっている．これより，共通話題の方が関連する論文が多く，対応するクラスタの生成が困難であったと考える．

表 2 において，緑色の話題は「翻訳」と「音声」に関する話題である．「翻訳」と「音声」も両実験に共通して発見されている話題であり，含まれている論文数が多いと考える．また，「翻訳」「外国語の音声翻訳」「音声翻訳」のように，実験協力者が話題をどの程度まで細かく判断するかによって，共通話題であるか特有話題であるかに違いが出ることも観測された．この時，実験協力者が話題の粒度を細かく捉える場合には論文数が少なくなる傾向にあった．従って，話題の粒度がクラスタの形成しやすさに影響を与えたと考える．

表 3：平均発見話題数

| | 単語集合レベルあり | 単語集合レベルなし |
|---|---|---|
| 共通話題 | 2.75 | 2.25 |
| 特有話題 | 1.125 | 1.125 |

表 4 に各レベルの有用度の平均をそれぞれ示す．クラスタレベルと単語レベルは単語集合レベルの有無で有用度の平均に差があまり見られないのに対し，文書レベルの有用度は違いが大きくなっている．これは，文書レベルと単語集合レベルにそれぞれ対応したパネルは，単語の指定を行う点で役割が重複しているためと考える．また，単語集合レベルの有用度は高く，単語集合レベルに対応したパネルでの単語指定が行いやすかったと考える．

表 4：各レベルの有用度

| | 単語集合レベルあり | 単語集合レベルなし |
|---|---|---|
| クラスタレベル | 4 | 4.375 |
| 文書レベル | 3.125 | 4.375 |
| 単語集合レベル | 4.5 | – |
| 単語レベル | 3.375 | 3.5 |

表 5：有用度の高い情報と協調

| | 単語集合<br>レベルあり | 単語集合<br>レベルなし |
|---|---|---|
| クラスタリング結果 | 4.75 | 4.625 |
| 重みが変更された単語 | 4.375 | 4.5 |
| 文書本文 | 3.25 | 4.75 |
| 指定文書の単語一覧と出現頻度 | 4.75 | – |
| 指定単語を含む文書一覧 | 4.125 | 4.25 |
| クラスタレベルでの表示変更 | 4.5 | 4.625 |
| 文書レベルでの表示変更 | – | 4.125 |
| 単語集合レベルでの表示変更 | 4 | – |
| 文書番号のハイライト | 4.5 | 4.875 |
| 指定単語のハイライト | 3.75 | 4.75 |

表 5 に有用度が高いと評価された情報と協調を示す．クラスタリング結果やクラスタレベルでの表示変更など，作業を行う際に必須となるものの有用度は非常に高くなっている．また，文書本文，指定単語のハイライトは単語集合レベルの有無により有用度に差が出ている．先述のように単語集合レベルなしの場合は文書レベルの文書本文で単語指定を行うため，これらの有用度が高くなったと考える．

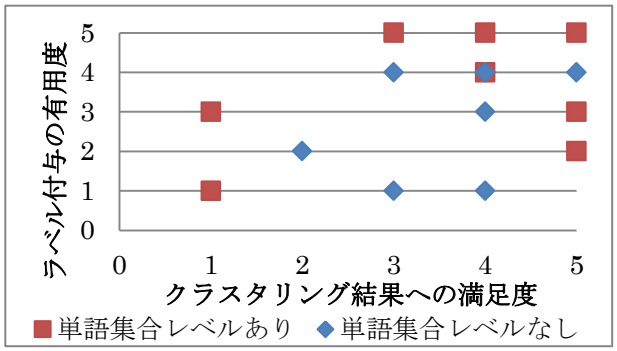


図 6：クラスタリング結果とラベル付与の関係

図 6 に，アンケートにより得られたクラスタリング結果への満足度とラベル付与機能の有用度の関係を示す．図より，ラベル付与が高評価の実験協力者はクラスタリング結果への満足度が高くなる傾向が見られる．これはラベル付与によって，発見した話題を見失うことなく，クラスタに関心のある論文が集まっているかどうかを効率よく確認することができたためと考える．ラベル付与を高評価していないが結果への満足度が高い実験協力者もいたが，その実験協力者は発見話題数が少なく，指定単語を含む文書番号がハイライトされるためラベル付与を高評価しなかったと述べていた．すなわち，比較的少ない話題であればラベル付与を行わなくても，指定単

語のハイライト機能によって話題を見失わず，クラスタリング結果の確認が可能であると考える．

表 6：視線データ

| | 単語集合レベルあり | | | | 単語集合レベルなし | | |
|---|---|---|---|---|---|---|---|
| | クラスタレベル | 文書レベル | 単語集合レベル | 単語レベル | クラスタレベル | 文書レベル | 単語レベル |
| 見た回数 | 633 | 559 | 585 | 103 | 139 | 186 | 39 |
| 平均時間[秒] | 0.75 | 1.03 | 1.21 | 0.3 | 1.42 | 2.12 | 0.68 |
| 総時間[秒] | 475.58 | 576.5 | 709.33 | 31.3 | 198.04 | 394.37 | 26.33 |
| 総クリック数 | 946 | 31 | 280 | 54 | 256 | 114 | 94 |
| 有用度 | 4 | 2 | 5 | 3 | 5 | 5 | 3 |

表 6 に，クラスタレベルを有用と判断した 2 名の実験協力者の視線データを分析した結果を示す．各パネルを AOI（Area of Interests）に設定し，パネル内に視線が入った回数･時間を測定している．また，当該実験協力者の gaze plot を図 7 に示す．左側が単語集合レベルのある実験協力者，右側が単語集合レベルのない実験協力者である．どちらもクラスタレベルを見た回数，総時間，総クリック数が多くなっており，作業中に多用した結果有用と判断したと考える．また，単語レベルの平均時間と総時間は比較的少ない．これは単語レベルの情報量が少なく，表示情報を確認するのにそれほど時間がかからなかったためと考える．

単語集合レベルがない実験協力者は，文書レベルの文書本文で単語指定を行うため，単語集合レベルがある場合よりも，文書レベルを利用する必要がある．表 6 を見ても，文書レベルの総時間と平均時間，総クリック数が，単語集合レベルがある場合と比較して多くなっていることがわかる．また，文書レベルの有用度は高いと回答されており，この場合も多用したことが評価につながったと考える．

図 7 において，両実験協力者とも，クラスタレベルから文書レベルといったように，隣接するパネルへの視線移動が多く見られた．このことから，実験協力者が文書の情報をクラスタレベルから単語レベルへと，レベル順に確認する傾向にあったと考える．

以上より，作業中の視認回数・時間やクリック回数と，有用度の間には関係が見られることから，実験協力者による有用度の評価は，作業の実態を反映したものと考える．

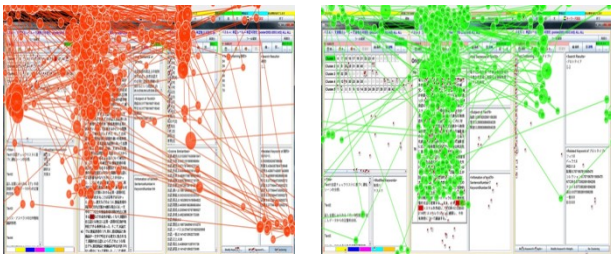

図 7：クラスタリング 1 回あたりの視線の動き（左：単語集号レベルあり，右：なし）

## 5. おわりに

本稿では，協調的マルチビューに基づくインタラクティブ文書クラスタリングシステムを提案した．ユーザ実験と視線データの分析を行い提案システムの有用性を示すと共に，ユーザにとって有用な情報や協調，提示する情報の違いが分析作業や実験結果に与える影響についても考察を行った．

本稿により得られた知見は，無駄な情報提示や協調の削減や，各情報の最適な提示方法の検討などといった，インタフェースの設計に貢献することが期待できる．

## 参考文献

[1]那須川 哲哉: テキストマイニングを使う技術/作る技術, 東京電機大学出版局, 2006.

[2]三宅 遼祐, 山田 誠二, 岡部 正幸, 高間康史: インタラクティブクラスタリングのためのマルチタッチインタフェースの提案, 第 25 回人工知能学会全国大会, 1J1-0S9-3, 2011.

[3]砂山 渡, 高間 康史, ダヌシカ ボレガラ, 西原陽子, 徳永 秀和, 串間 宗男, 松下 光範: テキストデータマイニングのための統合環境 TETDM の開発, 人工知能学会論文誌, Vol. 28, No. 1, pp. 1-12, 2013.

[4]市村 由美, 長谷川 隆明, 渡部 勇, 佐藤光弘: テキストマイニング-事例紹介, 人工知能学会誌, Vol.16, No.2, pp.192-200, 2001.

[5]J. C. Roberts: State of the art: Coordinated & multiple views in exploratory visualization, *International Conference on Coordinated and Multiple Views in Exploratory, pp.61-71, 2007.*

[6]T. Zhang, Q. Liao, L. Shi: Bridging the Gap of Network Management and Anomaly Detection through Interactive Visualization, *Pacific Visualization Symposium, pp.253-257, 2014.*

[7]C. Weaver: Building Highly-Coordinated Visualizations in Improvise, *IEEE Symposium on Information Visualization*, *pp. 159-166, 2004.*

[8]岡田 貴史, 石橋 融, 高間 康史: M2VSM を用いたテキストマイニングシステムの構築に関する考察, FSS2006, pp. 203-206, 2006.

[9]那須川 哲哉, 諸橋 正幸, 長尾 徹: テキストマイニング: 膨大な文書データの自動分析による知識発見, 情報処理学会, Vol. 40, No. 4, pp. 358-364, 1999.