2E4-NFC-01b-2

# TETDMを用いた協調的マルチビューによるインタラクティブ文書クラスタリングの提案

## Proposal of CMV-based Interactive Document Clustering System Developed on TETDM

高間 康史*1
Yasufumi Takama

利根川 拓馬*1
Takuma Tonegawa

*1 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

This paper propose an interactive document clustering system, which is designed based on the concept of CMV (cooperative multiple views). A prototype system is implemented on TETDM (Total Environment for Text Data Mining) because it can provide the mechanism of cooperation between modules. The proposed system classifies information to be presented into 4 levels: clusters, document, bag of words, and word, each of which is displayed with different view. This paper shows experimental results with test participants.

## 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．文書情報の探索的な分析を支援するインタラクティブクラスタリングは多種多様な情報の活用に有効であることが期待できる．しかし，インタラクティブクラスタリングシステムの開発においては「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」といった問題を検討する必要がある．本稿では協調的マルチビューのコンセプトに着目し，ユーザに提示すべき情報を 4 階層に分け，それぞれを適切なビューに表示する．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．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によってインタラクティブなクラスタリングを実現し，各文書に対するラベル付与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．提案システムの実装には，モジュール間の連動に関する実装が容易な TETDM（Total Environment for Text Data Mining）[砂山 14]を採用する．ユーザ実験により，提案システムの有用性を示す．

## 2. 関連研究

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．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として，Overview + Detail views は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詳細部分（detail view）を表示する．応用例として，Zhang らは Overview + Detail views により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[Zhang 14]．

一般的な協調メカニズムに Brushing と Navigational Slaving がある．Brushing は，あるビューで要素を選択した場合に，リンクされた他のビューにおいて対応あるいは関連する要素が同時にハイライトされる．Navigational Slaving ではユーザがあるビューでスクロールやクリック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． Weaver 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[Weaver 04]．

連絡先： 高間康史，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，ytakama@tmu.ac.jp

## 3. 提案システム

提案システムはクラスタリング結果についての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 種類に分け，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 に各レベルにおいて提示する情報及び可能な操作を示す．

表1. 各レベルにおいて提示する情報及び可能な操作

| レベル | 提示情報 | 操作 |
|---|---|---|
| クラスタ | ・クラスタリング結果<br>・文書タイトル<br>・単語の重み | ・文書選択<br>・クラスタ数変更<br>・文書ラベル付与 |
| 文書 | ・本文<br>・重要文<br>・重要単語<br>・文数，単語数など | |
| 単語集合 | ・単語一覧<br>（出現頻度）<br>・単語間類似度 | ・単語選択 |
| 単語 | ・文書一覧<br>・意味<br>・類似度の高い単語 | ・文書選択<br>・単語の重み調整<br>・クラスタリング実行 |

文書 $d_i = (w_{i1},...,w_{in})$ における単語の重み $w_{ij}$ は tf-idf 値を用いる．クラスタリングには k-means アルゴリズムを採用し，単語の重みによってユーザフィードバックを与えるアプローチを採用する[Okada 08]．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する．全単語の $k$ の初期値は 0 とし，1 ずつ増減可能である．文書間類似度は式(1)によって計算する．

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

提案システムでは Brushing と Navigational Slaving によるレベル間の協調を可能とする．ユーザが文書番号や単語をクリックにより選択することで，表示内容の変更や関連要素のハイライトなどがパネル間連動により実行され，文書情報や文書間の関係性を効率良く把握できる．図 1 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，図中左端から順にクラスタ，文書，単語集合，単語レベルに対応している．また，単語レベルのパネルを用いて単語の重み調整や再クラスタリング，クラスタレベルのパネルを用いてクラスタ数変更や文書へのラベル付与が可能である．ラベルはユーザが関心を持った文書に対して付与され，クラスタリング結果を調べる際に利用される．

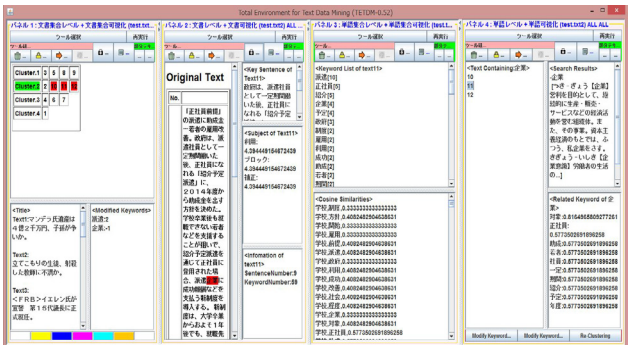

図 1. 提案システムのインタフェース

## 4. 評価実験

### 4.1 実験の概要

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が異なるシステムをそれぞれ 8 人ずつに利用してもらった．実験終了後には，発見した話題や提案システムの提示情報・機能などに関するアンケートに回答してもらった．

### 4.2 実験結果

表 2 に各レベルの有用度の平均を示す．有用度は1：低評価から 5：高評価の 5 段階で評価してもらった．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単語集合レベルありの場合に低くなっている．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，単語の指定を行う際の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，単語の選択などで活用されたと考える．

クラスタリング結果に対する満足度とラベル付与機能の有用度の関係を図 2 に示す．○×は単語集合レベルの有無を表す．ラベル付与が高評価の場合にクラスタリング結果に満足する傾向が見られるが，ラベル付与によって着目した話題に関する論文がグループ化されているかを効率よく確認できたためと考える．

表2. 各レベルの有用度

| レベル | 単語集合レベルあり | 単語集合レベルなし |
|---|---|---|
| クラスタ | 4 | 4.4 |
| 文書 | 3.1 | 4.4 |
| 単語集合 | 4.5 | – |
| 単語 | 3.4 | 3.5 |

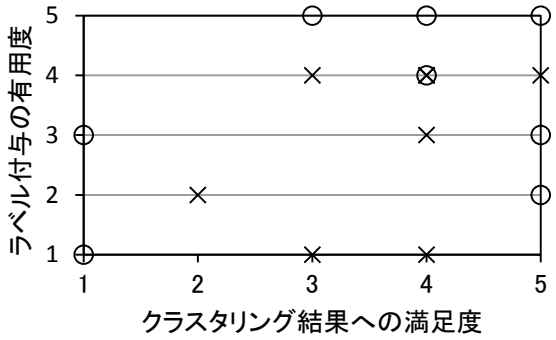


図2. クラスタリング結果とラベル付与に対する評価の関係

## 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，ユーザ実験により有用性を示した．本研究により得られた知見は，無駄な情報提示や協調の削減，各レベルでの最適な情報提示方法の検討といった，インタフェース設計に貢献することが期待できる．今後，クラスタリング処理時間の改善などを行い，大規模データセットを分析可能とすることで，よりユーザの意図を反映させたインタラクティブ文書クラスタリングシステムの実現が可能になると期待できる．

### 参考文献

[砂山 14] 砂山渡，高間康史，西原陽子，梶並知記，串間宗夫，徳永秀和：統合環境 TETDM を用いたマイニングツールの開発と利用の実践，人工知能学会論文誌，Vol.29, No.1, pp.100-112, 2014.

[Zhang 14] 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.

[Weaver 04] C. Weaver: Building Highly-Coordinated Visualizations in Improvise, IEEE Symposium on Information Visualization, pp. 159-166, 2004.

[Okada 08] Y. Takama, T. Ishibashi, T. Okada, Y. Horii: M2VSM: Extended VSM based on Meta Keyword and Its Application to Text Mining, Int'l J. of Computer Science and System Analysis, Vol. 2, No. 2, pp. 115-120, 2008.