F-043

計量学習を用いたユーザー意図学習システム

Learning User's Intention Using Metric Learning

門馬　道也†　　森永　聡†　　河村　大輔†
Michinari Momma　Satoshi Morinaga　Daisuke Komura

1. まえがき

本論文では、テキストクラスタリングのタスクにおいて、システムとのインタラクションを繰り返しながら、分析結果にユーザー意図を反映していくフレームワークを提案する。

ユーザーの領域知識や分析者の観点をデータマイニング結果に反映することは、非常に重要なことである。例えば、テキストクラスタリングでは、不要語/類義語リストの作成といった前処理や、クラスタ群をマップ表示する際の配置の工夫等の後処理といった方法によることが一般的であった。これらは、クラスタリング処理とは独立に、大きな工数を伴って手作業で行われることが多いため、前後処理を含めたトータルの分析コストという面で問題であった。

本研究では、クラスタリング結果に対するユーザーの指示から、意図を文書ベクトル空間の計量として学習し、それを用いて再クラスタリングする、ということを繰り返すというアプローチで、上記の問題を改善するシステムを提案する。ユーザーからの指示は、直感的な統合ＧＵＩを用いて簡単な操作で行うことができるので、トータルの分析コストが大きく低減する。

2. ユーザー意図学習システム

図１は提案するユーザー意図学習システムの全体像である。本システムは大きく、距離ベースのクラスタリングエンジン、統合ＧＵＩ、計量学習エンジンの３ユニットからなる。

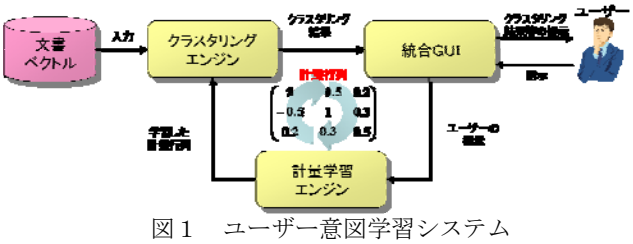

図１　ユーザー意図学習システム

クラスタリングエンジンは文書ベクトルの集合と、その空間の計量行列を入力として、文書群を近いもの同士にクラスタリングする。ユーザーは統合ＧＵＩを通じ、その結果や判断用の補助情報を見た上で、“二つのクラスタを結合せよ”等の指示をシステムに出し、計量学習エンジンは出された指示を実現するように計量行列を学習する。そして、それを用いて再度クラスタリングが行われ、結果がユーザーに提示される。このようなループを繰り返しながら、徐々にテキストクラスタリング結果へのユーザー意図の反映が進んでいき、最終的に分析者の領域知識や観点を反映したクラスタリング結果が得られる。

クラスタリングエンジンは、文書間の距離に基づくアルゴリズムであれば、k-means やグラフ分割等の通常のものでよいが、文書ベクトル空間の計量は入力された計量行列を用いて定義するものとする。この場合、計量行列中の大きな対角成分に対応する単語は、文書間の距離を計算する際に重要視されることになり、大きな非対角成分に対応する単語ペアは、文書ベクトルの空間で距離が小さく計算されることになる（類義語とみなされていることに相当）。ユーザー指示に基づき、計量行列をうまく学習することで、その意図に沿うように各単語の重要性や単語間の類義性が調節され、分析者の領域知識や観点がクラスタリング結果に反映されることとなる。

統合ＧＵＩは、クラスタリングの結果や、学習された計量行列の内容、さらにはユーザーが指示内容を検討するための補助情報をグラフィカルに表示する一方、ユーザーからの様々な指示を受け付け、計量学習エンジンに入力する。トータルの分析コスト削減のためには、本ユニットの使い勝手は極めて重要である。表示例や可能なユーザー指示の種類等、統合ＧＵＩの詳細は第３章で詳細に説明する。

計量学習エンジンは、ユーザーから入力された指示がなるべく実現されるように計量行列を学習する。入力された指示は、それぞれ計量行列の満たすべき制約条件に変換し、その下での、前回の計量行列からの更新量を目的関数とした非線形凸最適化問題として学習アルゴリズムを定式化している。紙面の関係上、詳細な表式は掲載できないが、基本的に、様々な制約条件をヒューリスティックスにより Information theoretic metric learning [1]の枠組みに当てはめ、その逐次アルゴリズム（Bregman 法[2]）を構成することにより効率化している。

3. 統合ＧＵＩを用いたユーザー指示

図２は、統合ＧＵＩの画面例である。画面中央にはクラスタリング結果の表示部、上部にはユーザーが指示種別や画面モードを指定するボタン群、画面下段にはユーザーが指示内容を検討するための補助情報表示部が配置されている。

結果表示部においては、各クラスタは円柱で表示され、その体積は所属する文書数、半径は所属文書ベクトルの（計量行列を使って測った）散らばり具合を反映している。円柱のラベル文字列は、クラスタ ID やその特徴語等を表し、クラスタ中心間の（計量行列を使って測った）距離の小さいクラスタ同士は、線で結んで表示される。

本システムにおいてユーザーが出せる指示は、特定のクラスタに対する「必要」「不要」「分割」や、クラスタのペアに対する「結合」「結線」「断線」、特定の単語に対する「必要」「不要」や、単語のペアに対する「結線」「断線」である。

ある特定のクラスタがうまくまとまっているとユーザーが判断した場合には、ユーザーはそのクラスタに「必要」と指示をだし、おかしな観点でまとまっていると判断した場合には「不要」と指示を出す。「分割」は、二つのクラ

†　日本電気株式会社共通基盤ソフトウェア研究所, NEC

スタをなすべきものが一つにまとまっている場合に指示し、クラスタペアの「結合」は逆の場合に指示する。クラスタペアの「結線」は、内容が近いと判断されるのに線で結ばれていない場合に指示し、「断線」はその逆の場合に指示する。

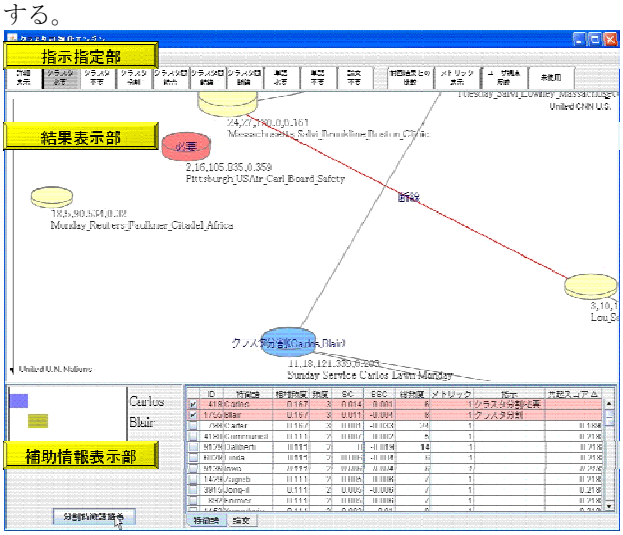

図２　統合ＧＵＩ画面例

上記のように「うまくまとまっているか」等をユーザーが効率よく判断するために、補助情報表示部では各クラスタの特徴語やその統計情報、原文検索機能等を提供している。特に「分割すべきか」の判断は難しいが、これを支援するために、特徴語同士の共起しにくさを指標にした「分割候補推薦機能」も提供している。

また、図３は画面モードを切り替えて、学習された計量行列をグラフィック表示した例である。大きなフォントの単語は対角成分が大きく、クラスタリングで重要視されているもの、単語間の結線は非対角成分が大きく、類義視されているものである。ユーザーはこの画面で、システムによるユーザー意図の学習結果を確認し、必要な場合には単語の「必要」「不要」「結線」「断線」等の指示をだす（一部は補助情報表示部からも可）。

本システムにおいては、ユーザー指示は全て、その種別と対象クラスタ/単語をマウス操作で指定することのみで実行可能である。例えば、図２はクラスタ「必要」や「分割」等を指示した状態である。このように、判断のための補助情報提供や直感的な操作性を統合ＧＵＩで実現することで、トータルな分析コストを大きく低減している。

4.分析例

図４は TDT pilot corpus[3]のうち、判定ラベルが”YES”[1]の記事を、本システムを用いて分析した結果である。被験者は事件ごとに別クラスタに分けるようにと言われて作業を行った。全体のループは３回繰り返され、全作業は４５分で終了した。なお、図２は計量学習を行う前（計量行列は単位行列）のクラスタリング結果である。

図２では Monday や Sunday など事件を特定できないクラスタが多いのに対し、図４ではほとんどのクラスタが事件を特定するものであり、ユーザー意図を反映したクラスタが形成されていることが分かる。また、図３は図４の結果を得た際の、学習された計量行列を表示したものであるが、Baghdad, Iraq, Americans など、アメリカ人がバグダッドで拘束された事件や、Bobby, Hall, North など、Bobby Wayne Hall という人物が北朝鮮で拘束された事件を特定するための単語が重要視され、お互いに結線されている。以上から、短期間で、計量行列の適切な学習、および、クラスタリング結果へのユーザー意図反映が行われたことが確認された。

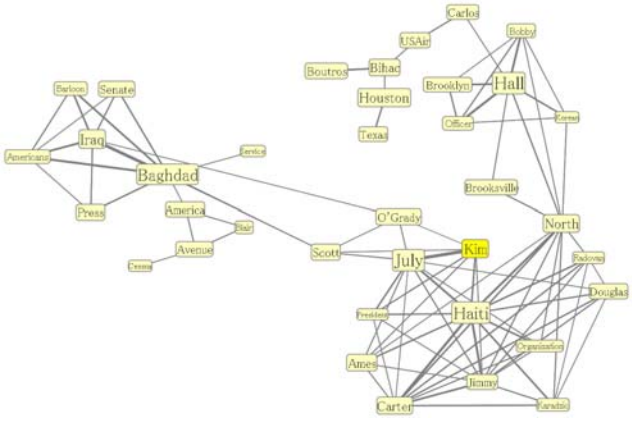

図３　学習された計量行列のグラフィック表示

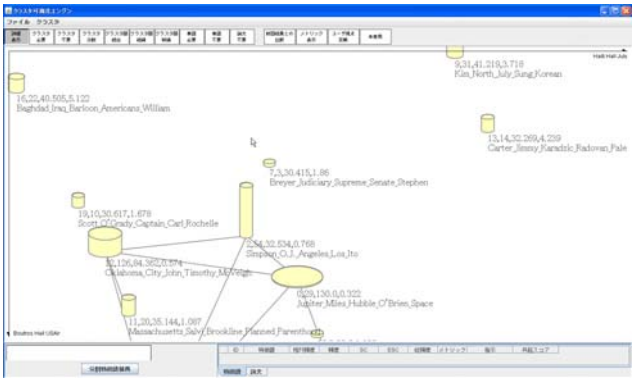

図４　本システムでの分析結果

5. まとめ

計量学習を用いて、テキストクラスタリング結果にユーザー意図を反映していくシステムを提案した。補助情報の提供や、直感的な操作性により、トータルな分析コストを大きく低減した。

参考文献

[1]　J.V. Davis, B. Kulis, P. Jain, S. Sra and I.S. Dhillon. Information-theoretic metric learning, ICML’07

[2]　Y. Censor, S.A. Zenios, Parallel optimization, Oxford University Press, 1997

[3] TDT Pilot corpus, http://projects.ldc.upenn.edu/TDT-Pilot/

[1] 文書のほぼ全体が１つの事件を述べている