1B1-2

編年型データ解析ツールの開発

An Analysis Tool for Chronological Data

赤石 美奈[*1] 伊藤 直之[*2] 箱石 大[*3] 石川 徹也[*3]
Mina AKAISHI Naoyuki ITOH Hiroshi HAKOISHI Tetsuya ISHIKAWA

[*1]東京大学大学院・工学系研究科, 科学技術振興機構さきがけ
School of Engineering, The University of Tokyo, PRESTO, Japan Science and Technology Agency

[*2]大日本印刷株式会社
Dai Nippon Printing Co., Ltd.

[*3]東京大学史料編纂所
Historiographical Institute, The University of Tokyo

This paper proposes a visual analysis tool to find relationships among chronological data and attributes. It helps users to verify a theory or to get new idea. The tool integrates a distribution map of data with views of frequencies and cooccurrences of attributes. They are interactively operating together. In this paper, we present a method of visualization and examples of data analysis.

1. はじめに

様々な分野において、時間属性を伴うテキストデータ[*1](政治、経済等のニュース、会議録、ブログ等の編年データ) が大量に蓄積されており、その有効活用が望まれている。そこで、本研究では、ユーザ自身が、それらのテキストデータに対してシステマティックな処理を行い、視覚的パターンを抽出し、詳細分析することで、有用な知識を導き出すことを支援することを目指す。

本論では、大量の編年型データを対象とし、そこに含まれるトピックや付与されたメタデータの経年変化を視覚化することにより、漠然と認識していたことを数値的に確認したり、新たな発見のための気づきを得ることを支援するための解析ツールを提案する。さらに、提案システムを膨大な歴史資料 (史料) に適用し、その有効性について報告する。

2. 編年型データの視覚化

2.1 属性データの関係性の俯瞰

大量のデータから、有用な知識を獲得する場合に、データの概要や属性の関係性がわかっている場合には、妥当なテキストマイニング、データマイニングの手法を適用することが考えられる。しかしながら、属性間にどのような関係があるのかわからない場合には、ユーザが試行錯誤しながら、関係を見つけていく作業が必要となる。

そこで、本研究においては、編年型データの解析の汎用的な手段を提供することを目的とし、(i) データ全体の俯瞰、(ii) 分析の着眼点の獲得、(iii) 詳細分析という解析プロセスデザインに基づき、データを視覚化する編年型データ解析ツール (CAT:Chronicle Analysis Tool) を開発した。

2.2 編年型データ解析ツール

本論で扱う編年型データは、テキストデータ、時間属性、その他の属性から構成される。編年型データ解析ツール (CAT) は、ユーザが、*scope*、*filter*、*viewpoint*、*attribute* の 4 つのパラメータを設定することで、情報の時系列分布状況を俯瞰し、解析の観点を見つけ、詳細分析に至る解析プロセスを支援するシステムである。*scope* は、扱うデータの時間属性に関する範囲を規定する。*filter* では、扱うデータに対す絞り込み条件を指定する。*viewpoint* は、対象データを捉える観点として、特定の属性と属性値を指定する。*attribute* は、対象データを分析する切り口として、興味のある属性の指定に用いる。つまり、CAT は、*scope* で規定された範囲の編年型データに対して、*filter* 機能でデータ集合を絞り込み、*viewpoint* にて指定した属性値を分析の着眼点として、着目する属性データの関係を視覚的に表示する機能を有する。

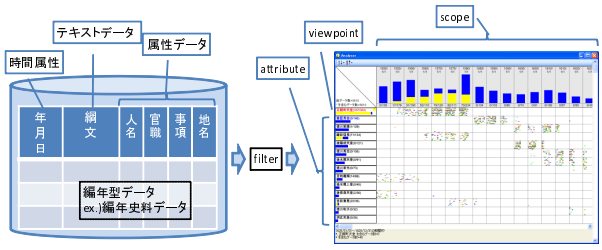

図 1: 編年型データ分析ツール.

図 1 にシステムの構成、及び、画面のハードコピーを示す。

横軸は時間を表し、縦軸は、*attribute* で指定された属性のデータが出現頻度順に並べられる。マトリクス部分には、該当する時間帯に、縦軸で示された属性値をもつレコードが、矩形アイコンとして並べられている。アイコン部にマウスを移動することで、ツール下部にテキスト内容や属性値などの詳細情報が表示される。

3. 有効性評価

3.1 評価方法

本システムの有効性を定性的に評価するために、編年型データを大量に保有する歴史学研究を対象とし、システムの有効性を検証する。

連絡先: 赤石美奈, 東京大学大学院・工学系研究科, 東京都文京区本郷 7-3-1, akaishi@ailab.t.u-tokyo.ac.jp

*1 本研究では時間属性を伴うテキストデータのことを「編年型データ」と総称する。

表 1: CAT を用いた分析プロセス.

[step1] 全体の俯瞰
scope として、対象データ全体を含む範囲を設定する。*filter* として、解析対象データを絞り込む条件を設定する。*viewpoint* は、設定しない。(全体像がわからないため、この時点では設定できない。) *attribute* には、興味のある属性を指定する。操作の結果として、*attribute* にて指定された属性値が頻度順に示され、その時間分布状況がマトリクス部に示され、データ全体の状況が俯瞰できる。
[step2] 分析の着眼点の獲得
表示されたデータに対して、高頻度や、高密度の時間分布部分がある属性データに着目し、これを分析の着眼点として、*viewpoint* を設定する。*attribute* には、興味のある属性を指定する。この結果、*viewpoint* で指定された属性データと共起する属性データの状況が表示される。
[step3] 詳細分析
step2 を繰返し、絞り込まれた対象データに対して、テキストの内容検討を行い、詳細を分析する。

3.2 対象データ

対象データとしては、分析結果の妥当性を評価するために、既に研究が進められている『復古記』[1] を取り上げることとした。『復古記』は、1867 年 10 月 14 日から 1869 年 6 月 12 日までの戊辰戦争の過程が記述されており、戊辰戦争研究において重要な編年史料である。正記と外記 (戦役/戦線別の記録) の二部構成からなり、正記は 2,485 件、外記は 3,336 件がデータ登録されている。本研究では、東京大学史料編纂所データベースの維新史料綱要データベースに納められている復古記のデータを利用した。

3.3 結果

CAT を用いた分析者の解析プロセス手順を表 1 に示す。さらに、表 1 に基づく分析プロセスの事例の一部と、それに対する歴史学研究者のコメントを表 2 に示す。

分析者の分析結果は、歴史学研究者によって、いずれも妥当であると評価され、本システムは、データを漫然と見ているだけではわからない、正確な知識の提供ができるツールであるとの評価を受けた。また、歴史学研究者にとっては、従来、史料を読んで漠然と認識していたことを数値的に確認できるということに意味があり、新たな研究テーマ発見のための気づきを与えられるといえる。興味深い研究課題を見つけ出すことをサポートできる可能性があるとの総評であった。しかしながら、今回の分析結果においては、歴史学研究者に対して新しい知見を提供するには至らなかった。この点は、今後の課題として残されている。

表 2: 『復古記』に対する分析事例.

	事例 1	事例 2	事例 3
	step1	step1	step1
s f v a	1867.10.1 -1868.10.31 『復古記』正記 - 事項	同左	同左
	step2	step2	step2
v a	事項:帰藩 事項	事項:大総督府 事項	事項:外交 官職
v a	事項:藩政改革 事項	事項:罷免 事項	官職:外国官副知事 人名
v a			人名=東久世通禧 官職
	step3	step3	step3
	帰藩が高頻度で出現し、帰藩と関連が深く出現する藩政改革が、帰藩の理由として挙げられる。	大総督府と、罷免、謹慎が高頻度で共起し、ここから罷免や謹慎が大総督府の重要な職務・権限に関わる事がうかがえる。	東久世通禧という人物が神奈川県裁判所総督、神奈川府知事、外国官副知事を歴任し、明治初期の外交で活躍していたと思われる。
歴史学研究者のコメント			
	“ 帰藩 ”というような大名の動向を相当重視した編纂がなされていたことを数値的に確認できたことが意外な結果であり、その“ 帰藩 ”の理由として上位にきたのが“ 藩政改革 ”であったことを確認できる。	大総督府の権限は非常に大きく、重要な機能として”罷免”権を行使しており、実際に”罷免”に関する多くの事項が記述されていることを確認できる。	歴任した役職は、当時の外交において非常に重要な役職である。一般的には、それほど有名ではないと思われる”東久世通禧”についての妥当な知識を得られている。

4. おわりに

編年型データ解析ツールは、編年型データを対象として、有用な知識を導き出すための視覚的解析ツールである。本論においては、歴史学で、ある程度、研究の進められている『復古記』を編年型データとして取り上げ、CAT を用いて、着眼点や分析の切り口を見つけ出し、知識を抽出するプロセスを示し、分析事例とその評価について述べた。分析結果として、専門家から見て、当たり前の事が抽出できたことは、システムの妥当性を示すものであると考える。

しかしながら、専門家にとって未知の知見を導きだすための支援が今後の課題として残されている。今回は、分析の着眼点を獲得する際に、密度や頻度の高いものに着目したが、共起度等、他の指標により、新しい着眼点を得られる仕組みを明らかにしたいと考えている。

また、他分野の編年型データや、人間の手作業では処理しきれない莫大な編年型史料データに適用することで、専門家が気づいていなかった知識を導き出す可能性について検討を進め、本システムの有効性について、さらに検証していきたいと考えている。

参考文献

[1] 東京帝国大学文学部史料編纂所：復古記　第 1 冊～第 15 冊（1929～1931）.