1H5-NFC-01b-3

統合環境 TETDM を用いた テキストマイニング初心者のスキル獲得支援

Acquisition of Text-Mining Skills for Beginners Using TETDM

中垣内 李菜 Rina Nakagochi　川本 佳代 Kayo Kawamoto　砂山 渡 Wataru Sunayama

広島市立大学大学院情報科学研究科
Graduate School of Information Sciences, Hiroshima City University

Only the very limited people with text mining skill can acquire significant knowledge using text mining tools effectively. TETDM assume not only experts of text mining or this software but also the beginners as users. However, beginners had to master how to use it by trial and error repeatedly until now. Hence, in this paper, we defined the skill to operate text mining tools smoothly and to analyze results appropriately as text-mining skill. Then we proposed a system which assists for beginners to use TETDM and showed the effect of the system by experiment.

1. はじめに

近年，安価で大容量の記憶装置や，高性能の演算処理装置の普及，ネットワーク環境が整備されたことにより，膨大な量のデータが取り扱われるようになり，「データマイニング」が注目されるようになった．テキストを対象としてデータマイニングを行うツール の一つに統合環境 TETDM[砂山 13]がある．統合環境であるTETDMは，テキストマイニング技術を提供している複数の処理ツールと，画面への出力を行う複数の可視化ツールを柔軟に組み合わせて，ひとつのソフトウェア内で動作させることが可能な統合環境である．TETDM が想定している利用者は，テキストマイニングやソフトウェア利用の熟練者だけではなく初心者も含まれる.これまでは TETDM に対面した初心者はどこからどのように学べばよいかわからないまま,解説を読み, 試行錯誤を繰り返しながら使い方を習得していた．本論文では，テキストマイニングや TETDM の初心者が無理なく導入から初級レベルまでを習得できるようになるためにチュートリアルを実装し，その効果を評価実験により示すとともに確認し，さらに，使用者の特性による効果の違いから改善すべき点を明らかにすることを目的とする．なお，初級レベルとは一通りの処理ツールと可視 化ツールの使い方に関する知識を習得し,それらを生か して各自の目的に応じたテキストマイニングを試みることができるレベルとする．また，本研究における初心者は TETDMを0から数回使ったことがあるか特定のツールのみを使用したことがある人とする．

2. 関連研究

2.1 統合環境 TETDM

TETDM は, 複数のテキストマイニング技術を柔軟に組み合わせて使用可能な統合環境を構築し, 社会的創造的活動を支援できる環境の提供を目指している. TETDM では, 学生や主婦など,PC は利用するがテキストマイニングという言葉を知らないユーザも想定しており, 単純かつ直感的に用いられ,利用価値がある環境の構築を目指している. TETDM に関して, これまで処理ツールや可視化ツールの開発に関する様々な研究[砂山 01] [山手 12][梶並 12]が行われてきたが，本研究では，TETDM を初めて使う人が, TETDM の機能をより容易に, より有効に使えるようにするために, TETDMを使ったテキストマイニング学習のための新しいインタフェースを提案する.

連絡先:中垣内李菜, 川本佳代, 砂山渡,広島市立大学大学院情報科学研究科システム工学専攻, 広島市安佐南区大塚東三丁目 4 番 1 号,{nakago,kayo,sunayama}@sys.info.hiroshima-cu.ac.jp

2.2 既存のテキストマイニングツール

すでにデータマイニングツールは数多く存在する．例えば，生物学の分野で，PolySearch[Cheng 08] や GeneWizard[Faro 11]や@Note[Lourenço 09]が開発されている．これらのツールは，研究者が膨大な量の科学・技術・医学雑誌からの需要な生物医学情報を取得することによる研究の促進を目指している．これらをはじめとして既存のデータマイニングツールの研究のほとんどは，対象のユーザが特定分野の専門家であることを前提としている．日本で，公開されているテキストマイニングツールには，DIAMining[DIAMining]，Text Mining Studio[Text Mining Studio]，TRUE TELLER[TRUE TELLER]，LanguageWare [LanguageWare]などがある．これらのソフトウェアは使用のために説明書やサイトで具体例の提示などが用意されているが，主にビジネスの現場にいる人をユーザとして想定しているため，利用するユーザには高いモチベーションを維持し各自でテキストマイニングの知識を得たり，経験を積むことが求められている． 本研究では, ユーザがテキストマイニングを実際に行いながら自然と知識と技術を得ることができるようなインタフェースを作成する.

3. チュートリアルシステム

3.1 チュートリアルシステムの概要

本チュートリアルシステムのインタフェースはチュートリアルウィンドウ(図 1)，課題の詳細ウィンドウ(図 2 左)，課題解答ウィンドウ(図 2 右)の 3 種類のウィンドウから構成される．本チュートリアルシステムはあらかじめ作成した課題のタイトルを記述した「課題リスト」と各課題の詳細な内容を記述した複数の「各課題テキスト」を入力とし，各課題を一覧できるチュートリアルウィンドウを作成する．ユーザはチュートリアルウィンドウから任意の課題を選択することにより，チュートリアルに挑戦する．なお，現在初級レベルの課題を実装しているが，「課題リスト」や「各課題テキスト」を追加・変更することにより簡単に中級・上級者用のチュートリアル課題を追加・修正が可能となっている．

3.2 チュートリアルシステムのインタフェース

(1) チュートリアルウィンドウ

チュートリアルウィンドウでは，上部に学習者のレベルや経験値，使用者がマウスのカーソルを合わせている課題のタイトルを表示し，下部に複数の宝箱を表示する．チュートリアルは内容に応じて8種類(MISSION0 から MISSION7)の 46 課題が用意されており，各宝箱が各課題に対応している．最初の 3 種類の MISSION では,テキストマイニングや TETDM に関する基礎知識を得るとともに,TETDM の基本操作を一通り学ぶことができる．残りの 5 種類の MISSION では，提供されている処理ツールと可視化ツールの使い方,処理ツールと可視化ツールを組み合わせた使い方について学ぶことができる．各課題の宝箱をマウスでクリックすると，課題の詳細ウィンドウと課題解答ウィンドウが表示される．

(2) 課題の詳細ウィンドウと課題解答ウィンドウ

課題の詳細ウィンドウは上部に課題の詳細を表示し，下部にページ送りボタン(back <, next >)，画像表示ボタン(figure)，ウィンドウを閉じるボタン(close)を配置した．各課題は以下のいずれかに対応する行動を行うことによりクリアとなる．

- 課題の詳細ウィンドウとともに課題解答ウィンドウが表示された場合，課題の詳細ウィンドウの記述されている説明を読んだ後，必要なツールを探し出してセットして答えを探し，解答ウィンドウにて答えを記入し正解する．
- 課題の詳細ウィンドウ内の文章中に[クリア条件]という記述があった場合，[クリア条件]として書かれている操作を行う．[クリア条件]の内容は指定されたツールをセットしたり指定されたボタンを押すものである．
- 以上の 2 つの条件のいずれにも当てはまらなかった場合，課題の詳細ウィンドウに記述されている説明を最後まで読む．

3.3 モチベーション維持のためのゲーム的要素

チュートリアルウィンドウ内で表示される各課題を示す宝箱は，マウスカーソルを合わせると開くようになっており，それぞれの課題をクリアすることで宝箱の中の宝が表示されるようになる．また，各 MISSION を達成するたびに次の MISSION の宝箱が選択可能になる．また，各課題をクリアするごとに経験値が入り，獲得した経験値による学習者のレベルを SKILL LEVEL として表示した．ゲーム的要素として学習者がモチベーションを維持しながら多くの知識やスキルを獲得できるためにこれらを実装した．

図 1: チュートリアルウィンドウの表示例

図 2: 課題の詳細ウィンドウ(左)と課題解答ウィンドウ(右)の表示例

4. チュートリアルシステムの評価実験

4.1 実験目的

提案したチュートリアルシステムが TETDM 使用の支援に有効であるか，また，使いやすい物であるか，さらにどのような特性をもつ利用者にとって効果的かを明確にすることを目的とした．

4.2 実験方法

被験者は成人の男女 20 名である．被験者には表 1 に示す手順に従ってチュートリアルを使用してもらった．評価は，手順 3 のチュートリアルの前後に行われる，制限時間各 30 分のテストの得点を比較することで行った．このテストは，TETDM 内のツールを利用することで解答できるテキストマイニングの初歩的な問題で構成されている．ただし，チュートリアルの内容はツールの使い方を説明するのみであるが，解答には適切なツールの選択と操作が必要となるため，チュートリアルの内容がテストの解答に直結するわけではない．

また，被験者の特性と正答率や課題の解答にかかった時間との関係を調べるための特性調査アンケートと，システムの「使いやすさ」について主観的評価を得るための事後アンケートに答えてもらった．

表 1: チュートリアルシステム評価実験の手順

	実験手順詳細	想定所要時間(分)
1.	MISSION0, 1, 2 全 21 課題の遂行 課題：一通りの操作方法の修得	30
2.	事前テストへの解答 テキストマイニング初級問題への解答	30
3.	MISSION3 から 7 全 25 課題の遂行 課題：一通りのツールの使い方の習得	40–60
4.	事後テストへの解答 事前テストと同レベルの問題への解答	30

4.3 実験結果と考察

(1) 事前/事後テストの得点

事前テストと事後テストの平均得点(事前テストと事後テストはそれぞれ 20 点満点)について t 検定の結果を表 2 に，各被験者の平均得点を図 3 に示す．表 2 より，事前テストと事後テストの平均得点を比較したところ，事後テストの方が有意に高くなった．また，図 3 より，20 人中 19 人の被験者が事前テストよりも事後テストの方が良い点をとった．事後テストよりも事前テストの方が良い点を取ってしまった被験者 α のテストの詳細情報を表 3 に示す．表 3 より，被験者 α は不正解数が 3 減り，無記入数が 4 増えていることから，チュートリアルによって得た知識をもとにじっくりテストに取り組んだため時間が足らず，正解数が減ってしまったと考えられる．このことから，本チュートリアルは一通りの処理ツールと可視化ツールの使い方に関する知識を習得させ，それらを生かして各自の目的に応じたテキストマイニングを試みるスキルを身につけさせる上で，一定の効果があったといえる．しかし一方で，事後テストの点数は十分に高いと言える点数ではなかった．これは，チュートリアルでは使用するツールが指定された上で，その処理を試す内容となっていたのに対して，テストでは 30 以上の処理ツールと 30 以上の可視化ツールの中から，適切なツールを選ぶ必要があったことが影響したと考えられる．よって，目的に応じたツールの選択を支援する方法を整備することは，今後の重要な課題である．

表 2: 事前/事後テストの平均得点

	事前テスト	事後テスト
被験者数 (人)	20	20
平均(点)	9.15	13.45
検定結果	t(19)=6.28	p<.001

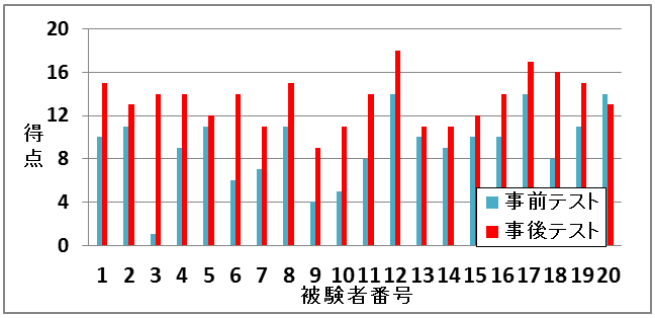

図 3: 各被験者の平均得点の推移

表 3: 被験者 α の事前/事後テストの詳細

	事前テスト	事後テスト
正解数	14	13
不正解数	4	1
無記入数	2	6
所要時間(sec.)	1800	1800

(2) チュートリアルおよび事前/事後テストの平均所要時間

チュートリアルの平均所要時間を表 4 に，事前/事後テストの平均所要時間を表 5 に示す．表 4 より，被験者らは MISSION0 から 7 の全 46 のチュートリアル課題を，平均約 62 分(標準偏差=約 23 分)でこなすことができ，予め設定した所要時間 90 分以内に，18 名が全課題を，2 名が 42 課題をクリアすることができた．よって，初心者向けに用意したチュートリアルの難易度は適切で，TETDM の使用方法の学習の効果が期待できる．また，表 5 より，事前テストと事後テストの平均所要時間に有意差はなく，制限時間 30 分以内にすべての問題を解けた被験者はいなかった．事前テストや事後テストにおいては，テスト中にテストを開始する前に達成したチュートリアルに再挑戦することを許可していたため，テストに解答するためにチュートリアルに再挑戦する被験者が多く見られた．そのため制限時間内にすべてのテスト問題に解答することができなかったと考えられるが，チュートリアルを何度もこなして慣れていくことにより短時間で解答できるようになると考えられる．

表 4: チュートリアル(全 46 課題)の平均所要時間

	チュートリアル
被験者数 (人)	20
平均(sec.)	3729.68

表 5: 事前/事後テストの平均所要時間

	事前テスト	事後テスト
被験者数 (人)	20	20
平均(sec.)	1795.50	1762.55
検定結果	t(19)=1.40 p = n.s.	

(3) 学習者特性と事前/事後テストの得点の関係

表 6 に信頼度の高かった学習者特性と事前/事後テストの得点の関係を示す．表 6 より，「TETDM を使ったことがある」学習者は 8 人で「事後テストの正解数 10 問以上」の学習者は 8 人，それ以外が 0 人で，二項検定を行った結果有意であった(両側検定: p=0.0078, p<.01)．TETDM を実験実施前までに使ったことのある学習者のすべてが事後テストで 10 点以上を獲得できたことから，TETDM 使用の経験があると，チュートリアルによる学習の効果が出やすいことがわかった．これまでは，試行錯誤により TETDM によるテキストマイニングを学習するしかなかったが，チュートリアルにより学習事項が明確になり学習の成果がはっきりでたと考えられる．このことから TETDM を使いこなすにはある程度経験することが必要であると考えられ，そこで TETDM を使ったことのない学習者に対してはその代わりとなる十分なチュートリアル課題を提示すると効果的であると考えられる．

表 6: 学習者特性と事後テストの得点の関係

ルール		出現頻度	特性出現頻度	成果出現頻度	信頼度(%)
特性	学習成果				
TETDM を使ったことがある	事後テスト 10 点以上	8	8	16	100

(4) 学習者特性とチュートリアルにかかった時間の関係

表 7, 表 8 に信頼度の高かった学習者特性とチュートリアルにかかった時間の関係を示す．表 7, 8 中の項目についてそれぞれ直接確率計算を行ったところ有意傾向が見られた(上から両側検定: p=0.0194, p<.05, p=0.0623, .05<p<.10)．プログラミングが得意な人や文章を書くときに一文の長さ

が長過ぎないように注意する人はチュートリアルを比較的短時間でこなすことができることがわかった．単純にコンピュータ使用に慣れている人が短時間でチュートリアルをこなすことが可能と考えられるため，コンピュータ使用に慣れていない学習者向けに専門用語をしっかり解説することや，わかりやすいインタフェース設計の徹底は重要であると考えられる．また，論理的な考え方をする人，日頃から文を書くときに読みやすさにこだわる人は文章を読むのに慣れていると思われるため，チュートリアルを効率よく学習できたと考えられる．現時点では，本論文にて作成したチュートリアルは，一つの課題に内容を詰め込みすぎない，長過ぎない文章を心がけて作成しているが，画像がないために文章を読み慣れない人は理解しがたい．よってチュートリアルの説明が難しいと思われる箇所には画像を提示できるようにするなどの工夫が必要であると考えられる．

表 7: プログラミングに関する学習者特性とチュートリアルにかかった時間の関係

ルール		出現頻度	特性出現頻度	成果出現頻度	信頼度(%)
特性	学習成果				
プログラミングが得意	チュートリアル 3000 秒以下で達成	6	8	8	75.0
プログラミングが苦手	チュートリアル 3000 秒以上で達成	10	12	12	83.3

表 8: 文章の記述に関する学習者特性とチュートリアルにかかった時間の関係

ルール		出現頻度	特性出現頻度	成果出現頻度	信頼度(%)
特性	学習成果				
文章を書くときに一文の長さに注意する	チュートリアル 3000 秒以下で達成	5	7	8	71.4
文章を書くときに一文の長さに注意しない	チュートリアル 3000 秒以上で達成	10	13	12	76.9

(5) 事後アンケート

実験終了後に，被験者に「チュートリアルを使うことによって，TETDM の使い方を理解できましたか？」というアンケートを行ったところ，「理解できた」と回答した人数と，「理解できなかった」と回答した人数は，18 :2 であった．その結果について，二項検定を行った結果有意傾向がみられた．また，「事前/事後テストは簡単でしたか？」というアンケートを行ったところ，事前テストと事後テストそれぞれについて「簡単だった」と回答した人数と，「難しかった」と回答した人数は，5:15 と 11:9 であった．その結果について，それぞれ二項検定を行った結果，事前テストでのみ有意傾向がみられた．よって，事前テストが難しいと感じる被験者が明らかに多かったのに対し，チュートリアルを行うことにより事前テストと同じレベルである事後テストを難しいと感じる被験者が少なくなったことがわかった．これは，各 MISSION のチュートリアルを順にこなすことにより TETDM 自体の操作方法やツールの使い方を無理なく習得できたためであると考えられる．一方で事後テストでも難しいと感じる被験者が多かったのはチュートリアルの内容が多く，実験に用いた数時間だけでは習得しきれなかったためと考えられる．

5. 結論

テキストマイニングや TETDM の初心者が無理なく導入から初級レベルまでを習得できるようになるためのチュートリアルを実装した．さらに，評価実験を行うことによりその効果を検証した．その結果，本チュートリアルシステムは，限られた時間内で一通りの処理ツールと可視化ツールの使い方に関する知識を習得させ，それらを生かして各自の目的に応じたテキストマイニングを試みるスキルを身につけさせる上で有効であることが明らかになった．

今後は，TETDM やコンピュータの使用しなれていないおよび文章を読み書きする経験の少ない使用者でも効果を出せるようなチュートリアルシステムを提案するとともに，さらに高度な中級・上級課題の実装して本システムの実用化を目指したい．

参考文献

[砂山 13] 砂山渡, 高間康史, 西原陽子, 徳永秀和, 串間宗夫, 阿部秀尚, 梶並知記: 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, 28(1),1-12, (2013)

[砂山 01] 砂山渡, 谷内田正彦: 展望台システムによる複数文書の要約と Web ページ集合への適用, 一般社団法人情報処理学会, 2001(86), 57-62, (2001)

[山手 12] 山手砂都美, 砂山渡: 文章の話の組み立てと展開速度による段落間関係の評価, 第 26 回人工知能学会全国大会, 3K2-NFC-3-4, (2012)

[梶並 12] 梶並知記: TETDM を用いた関連 Tweet 探索の一手法, 第 26 回人工知能学会全国大会, 3K2-NFC-3-7, (2012)

[Cheng 08] D. Cheng, C. Knox, N. Young, P. Stothard, S. Damaraju, DS. Wishart, PolySearch: a web-based text mining system forextracting relationships between human diseases,genes, mutations, drugs and metabolites, Nucleic Acids Research, Vol. 36, Web Server issue W399–W405, (2008)

[Faro 11] A. Faro, D. Giordano, C. Spampinato, Combining literature textmining with microarray data: advances for system biology modeling, Briefings In Bioinformatics. Vol. 13. No. 1. pp.61-82, (2011)

[Lourenço 09] A. Lourenço, R. Carreira, S. Carneiro, P. Maia, D. Glez-Peña, F. Fdez- Riverola, EC. Ferreira, I. Rocha, M. Rocha, @Note: A workbench for Biomedical Text Mining, Journal of Biomedical Informatics,Vol. 42. pp.710–720, (2009)

[DIAMining] 三菱電機インフォメーションシステムズ株式会社, “DIAMining”, http://www.mdis.co.jp/products/diamining/, (2014-03-10 アクセス)

[Text Mining Studio] NTT DATA Mathematical Systems, Inc., “Text Mining Studio”, http://www.msi.co.jp/tmstudio/, (2014-03-10 アクセス)

[TRUE TELLER] 野村総合研究所, “TRUE TELLER”, http://www.trueteller.net/, (2014-03-10 アクセス)

[LanguageWare] IBM, “LanguageWare”, IBM - United States, http://www-01.ibm.com /software/globalization/topics/languageware/, (2014-03-10 アクセス)