4J1-5

利用者の評価基準に合致した文章推薦システムの構築
感想データベースを応用した作品評価フォーマットの構築についての検討

秦野 智博
Hatano Tomohiro

阿部 明典
Abe Akinori

千葉大学大学院　人文社会科学研究科　阿部明典研究室

Abstruct:This study intend to construct a kind of "text recommendation system" that can find text coincide with personal preference. The former recommendation systems had some problems, typically "discordance of preference standard among user and system". To solve these issues, building "text analysis databases" and "recommendation sentence generator" now in progress. Moreover, building deatabase from text inpressions and reviews now investigated.

キーワード:テキストマイニング,検索システム,推薦システム,文章生成,文章解析
Keywords:text mining, searching system, recommendation system, sentence generation, text analysis

1. はじめに

現在の社会では，我々は膨大な分量のテキスト(文章)にアクセスすることが可能である．

書籍や雑誌といった出版物のみならず，インターネット上でアクセス可能なテキストは連続的に増加を続けている．加えてテキストへのアクセス方法も多様化しており，電子書籍の登場によって電子機器を介してインターネット上のテキストだけでなく出版物へのアクセスも可能になった．

このように，個人がアクセス可能なテキスト量が爆発的に増大していく一方で，読み手の目的に合ったテキストを探し出すのは困難になり続けている．

テキストを読む目的は，特定の事項に対する知識や意見が掲載されを抽出して役立てる「実用的」な読み方(主にニュースや論文，インタビュー記事，論説などが対象となる)や，何らかの表現を作品として受容する「娯楽的」な読み方(主に文芸や小説などが対象となる)など複数の形式に分類でき，実際にはそれらが併用されることもある．

どのような目的で読む場合でも，読み手はテキストを目に付いた順に無差別に読むわけではなく，テキストを読む段階以前でテキストを選別する必要がある．すなわち，あるテキストから目的に合う情報を得ることができるかどうかを，そのテキスト自体を読むことなく(厳密に言えば，テキスト自体を読むよりも短時間かつ労力の少ない方法で)判断しなければならない．

テキストに対する判断基準とは，読み手の嗜好と言い換えることも可能である．

ニュースであれば政治，経済，スポーツ，芸能といった分野ごとの選好に加え対象となる人物，団体，事件などへの興味関心が主要な嗜好だが，文章表現そのものも選好の対象となりえる．事実関係の表現や事件および当事者への書き手のスタンス，さらには文章の長さや語彙，用語の専門性など様々な要素を判断基準として挙げることが可能である．

我々は多くの場合，「検索」と「推薦」という方法を用いてテキストを選別している．

「検索」には機械検索の他，ジャンルごとの出版物一覧などから直接探す方法も含まれるが，ここでは機械検索に限定する．

連絡先:秦野 智博，千葉大学大学院 人文社会科学研究科 阿部研究室所属，千葉県印旛郡酒々井町下岩橋 448 番地 4
〒285-0907 電話 043-496-9776

多くの機械検索では入力キーワードがタイトルやジャンルに含まれているかを判定するのみなのでタイトルの絞り込みには有効だが内容については殆どわからない．そのため，具体的な推薦を併用してテキストへの判断を行うことが多い．

ここでいう「推薦」は，レビューなどの内容評価と機械的なレコメンドの両方を指す．

機械的なレコメンドはデータマイニング手法を応用したものであり，例としては，オンラインショッピングサービスなどを利用して商品の購入や商品情報を閲覧する際にみられる「類似性の高い商品の紹介」が挙げられる．

例のようなケースでは，利用者全体の行動履歴や商品のタグ情報，商品への評価などのデータを用いて推薦を行っている．この手法には広範な分野の大量の対象から比較的傾向の類似したものを選抜できるという長所が存在するが，テキストの内容ではなく外部情報に基づいた推薦であるため，具体的な類似の方向性や度合いを測りにくい点が短所である．

内容評価に基づく推薦とは，新聞や雑誌の書評欄，ウェブ上の個人によるレビュー投稿，直接の知人による評価（いわゆる「口コミ」）などを指すものである．

これらは実際にテキストにアクセスした人が行う内容評価であり，データマイニング的手法と比較すると，内容に関する情報がより詳細に得られるが，テキストの理解と解釈というプロセスが加わるために必要な労力が多く，評価できる作品数が少なくならざるを得ない．また評価対象が評価者にとって興味のある分野に限られてしまう点も差異といえる．

推薦や検索にはそれぞれの利点と欠点が存在するが，共通の課題として「評価者の評価基準と利用者の基準の不一致」が挙げられる．

既に述べたようにテキストの評価基準には複数の種類があるが，どの基準を用いる（重視する）か，用いない（重視しない）かも個人によって大きく異なる．特に娯楽作品の場合これは顕著で，ストーリーの展開や作品中の謎についての情報を公開してしまう，いわゆる「ネタバレ」情報はしばしば問題になる．単なる是非だけでなく，そもそも何をもって「ネタバレ」とするか人により見解が一致しないことも多く，利用者個人に一致する評価基準を用いた作品評価が理想的には求められる．

個人の評価基準(嗜好)に合致したシステムとはどのようなものであるべきか，どのような方法によって実現可能なのかが本研究の課題である．

2. 課題設定と構築過程

前段階で提示された課題は「個人の嗜好(テキスト評価基準)に合致した評価方法の構築」である．

この課題を解決するための具体的方法として，現在は既存テキストの解析による「データベース構築」，利用者の要求する基準に応じて適切に構築された「作品推薦文の生成，提示による推薦」という方法で解決できると考えた([難波 2006],[原田 2011],[松浦 2012]より参考)．

テキストの「データベース構築」とは，特定のテキストを評価要素となり得る様々な観点から解析，集積することで作品情報の多面的なデータベースを構築することを目指したものである．ここでは，現在実行している方法である，作品そのものの解析によるデータベース構築について述べる．

「作品推薦文の生成，提示による推薦」とは，構築されたテキストデータベースから利用者の要求する評価視点に基づき新たな文章を生成することである．生成された文章にはそのテキストの特徴や利用者にとっての適切性に関する分析が含まれ，最終的に利用者が対象テキストにアクセスするかどうかの判断材料となるものである．

さらに，作品データベースの構築について，新たに検討中の方法である「感想データベースの活用」についても検討を行う．

2.1 テキスト解析とデータベース構築

インターネット上から解析対象となるテキストを取得した．

主要な対象ウェブサイトは『青空文庫』やニュースサイトである．

作品の内容抽出に関しては，文章統計解析ソフトウェア『KH-Coder』を用いた数理的内容解析と，人力による内容解析の両方を用いた．なお，取得したデータの一部には解析を容易にするため，ルビの削除など内容改変を含まない加工を施している．

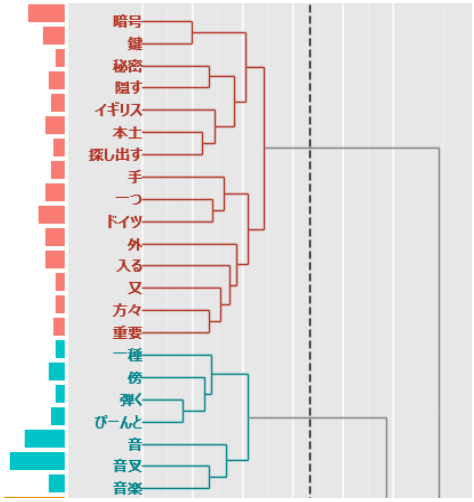
『KH-Coder』を用いた解析では，主に「抽出語リストアップ」と「階層的クラスター分析」を用いた．

「抽出語リストアップ」とはテキスト中の語を全て品詞分類し，更に品詞ごとに出現頻度の高い順にリストアップしたものである．

「階層的クラスター分析」とは共起関係にある語とその共起度合いを樹形図で示したものである．以下の[図 1][図 2]に実際の解析例から抜粋した図を貼付した．

名詞		サ変名詞		形容動詞		固有名詞	
メン	23	話	19	駄目	5	鬼ヶ島	1
音叉	21	安心	5	非常	5		
白木	19	発見	5	重要	4		
暗号	14	仕事	4	徹底的	4		
城塞	14	突破	4	妙	3		
音盤	13	失望	3	名誉	3		
蓄音機	13	出発	3	有名	3		
お嬢さん	8	振動	3	確実	2		
一つ	7	宝探し	3	豪華	2		
侯爵	7	用意	3	充分	2		
日本人	7	ダンス	2	必要	2		
本土	7	案内	2	無駄	2		
音楽	6	移動	2	たくみ	1		
部屋	6	廻転	2	はるか	1		
ステッキ	5	期待	2	ふしぎ	1		
銃声	5	研究	2	安心	1		

[図 1]抽出語リストアップ例の抜粋



[図 2] 階層的クラスター分析例の抜粋

手動による内容解析については，文の長さや使用語彙といった文体特徴の抽出に加え，作品内の特定の単語と全ての文についてラベリングを行った．

単語(品詞に限定されず，複合語も含む)レベルのラベリングについては，作品の登場人物のような重要度の高いものについてラベリングを行った．

[登場人物に対するラベル]

「主人公」:作中の事件に大きく関わり，語り手あるいは最も多く描写される人物を指す

「協力者」:例えば主人公の友人のように主人公の目的達成に積極的に協力する，あるいはその味方となる存在

「妨害者」:主人公の行動や目的を作為などで妨害する登場人物や、目的の達成を妨げる障害物を指す

「一般人」:事件の解決には直接関係しない登場人物

[事物に関するラベル]

「舞台」:作品の舞台

「事件」:刑事的な意味に限らず，作中で発生し主に主人公を巻き込むイベント全般

以下は主に文に付与されるラベルであり，人物や事物へのラベルと併用される．

「描写」:登場人物や舞台，事件の特徴や詳細を説明する要素

「行動」:基本的に人物のとった行動のうち，特に事件の進行や解決に寄与するものを指す

「心理」:登場人物が抱いたり表現した感情などを指す要素

個々の単語は同一のラベルを添付されることによって同一のグループと看做されることになる．特に小説作品の場合，作中で複数の呼称が使われることがあるため，人物の同定が必要となる．

また，「行動」や「心理」といったラベルとその主体および対象の関連付けも必要となる．

次に文レベルのラベリングについて述べる．こちらに関しては作中の全ての文単位(ただし台詞は全体で一文とした)で内容を解析し、内容に応じたラベリングを行った．また単語へのラベリングと同様、ラベルの名称は作品間での使用を考慮し，一般的・客観的な名称とした．

ここで理解促進のため，実際にデータベースを作成した小説作品の一つである『暗号音盤事件』を単語ラベリングの例として紹介する．なお，解析対象テキストは『青空文庫』より取得し，ルビの削除といった解析のための加工を施している．

この物語に登場する人物の一人で，主人公の友人として事件を共同で解決する人物「白木」は作中では本名の「白木豹二」，本人の特徴を表す複合語「勇猛密偵」，潜入先で名乗った仮名「二俵伯爵」と表記されている．これらは全て同じ人物の「白木」(作中で最も多く使われる呼称)を表すため，その全てに「協力者」のラベルを添付した．また，「白木」の人物像を表す表現を作中から抜き出し，それらに「描写」のラベルを添付すると共に「白木」の属する「協力者」とその性質を結びつけた．具体的には，「有名な」「勇猛な」「突拍子も無い」「我儘な」「社交的な」「人気のある」という単語がリストアップされた(一部、語形や仮名遣いの調整を行っている)．このような形で全ての登場人物及び舞台，事件についてそれぞれ対応するラベルとの関係性を整理した．

[文に対するラベル]

「種別」:文の形式が地の文か台詞かを区別する．

「分類」:文自体が作品中で果たす役割を示す．これには下位分類として「進行」，「説明」，「動作」，「心理」が存在する．

「進行」:物語の進行自体に関わる要素であることを示す．何らかの事件そのものや，事件を進展させる特定の出来事を指す．

「説明」:単語ラベリングにおける「描写」に相当する．事物の詳細や特定の事実に関する情報を登場人物や読者に伝える役割を有している．

「動作」:登場人物が行った何らかの動作を指す．単語へのラベルの「行動」とは異なり，事件自体への関連は問われない．

「心理」:いずれかの登場人物が抱いたり表現した感情や思考などを指す．

以下のラベルは、それぞれの対象が存在する場合のみ付けられる．なお、登場人物などを入力する場合は文中の表現ではなく単語ラベルのほうを利用する．

「主体」:ラベルはその文章の動作や感情の主体を指す．

「対象」:はその文章の動作や感情の対象，客体を指す．

「核心」:特殊なラベルで，基本的に物語の最終盤まで明かされない，明かされるべきでない情報(いわゆる「ネタバレ項目」)である文にのみ付加される．ここには特定の単語ではなく核心性の度合いを記入する．

文ラベリングの例として，[図 3]に実際の解析例からの抜粋を掲載した．

文	分類	主体	対
1国際都市			
私たちは、暫くの間リスボンに滞在することになった。	進行	主人公	
私の連れというのは、例の有名な勇猛密偵の白木豹二のことだ。	説明	仲間	
リスボンは、ポルトガルの首都だ。	説明	舞台a	
そのころリスボンは、欧州における唯一つの国際都市の観があった。	説明	舞台a	
この国は英米側に立つのでもなく、日本、ドイツ、イタリヤの枢軸国側	説明	舞台a	
だから、リスボンの町は、いわゆる呉越同舟というやつで、ドイツ人や	説明	舞台a	
だから私たちも、ここにいる間は別に中国人やベトナム人を装う必要か	説明	舞台a,主人公,仲間	
私は、久方振りのこうした安楽した気持におちついたので、願わくば、	心理	主人公	
その贅沢心を、或る日白木豹二が、一撃のもとに打ち壊してしまった	進行	仲間	
個はその前夜から宿を明け放しであったが、正午ごろになって、ふら	進行	仲間	

[図 3] 文ラベリング例の抜粋

このようにして作品解析を行い，テキストの特徴や「魅力」に関する情報データベースの構築を行った．

2.2 推薦文生成システム

データベース化された作品情報を基に，利用者側からの要求に応じて作品情報を選択し，作品推薦文を作成するシステムである．

推薦文の本体はテンプレートである．テンプレートは定型部分と空白部分から成り立っており，空白部分に利用者の求める情報を配置することで文章を作成，推薦文として提示する．

テンプレートの構成も複数の要素によって変動する．一つ目の要素は作品のジャンルである．現状では作品ジャンルと作品は固定的な対応関係にあり，利用者による変更はできない．

二つ目の要素が「利用者が重視する内容描写」である．現在のフォーマットでは，「人物描写の重視」，「舞台描写の重視」，「事件描写の重視」の中から利用者が重視する項目を選択する方式で構築している．それぞれの重視項目にあわせ推薦文を生成させる．

三つ目の要素は「利用者が重視する内容以外の描写」である．

具体的には，作品の文体特徴や描写の特徴にあたる部分であり，これについては特に作者間での変動が大きい要素であるが，「テンポの良さ」や「会話と描写文の比率」などの項目を設け利用者に制御させる方式で

四つ目にして重要な要素が「核心への言及」，つまりネタバレに関する項目である．前述の通り，作品内の要素が有するネタバレ度合いは「核心」のラベリングによって管理されている．この要素を利用者が操作する場合、ネタバレの程度を選択することになる．現在，このネタバレの程度については三段階で管理されており，利用者の望まないネタバレ情報が推薦文に盛り込まれることを防いでいる．

2.3 感想データベースの活用

「感想データベースの活用」とは，テキストに対する評価や感想自体を集積，解析しテキストのみの解析では得られない情報について収集する試みである．

小説を例としても，展開のアウトラインとしての「ストーリー」を抽出することは小説本文よりも小説に対する評価や感想からの方が容易に可能である．また，小説が対象とする読者(どのような嗜好を持つ者に適するか)に関する情報も同様である．

現在，書籍や新聞，インターネット上などから「書評」や「レビュー」，「感想」といったテキストに対する評価・感想文の収集を進めており，これらを用いた作品評価データベースの構築について検討を進めている．

3. 結果と考察

3.1 生成文章の例

実際に生成された文章の例を以下に掲載する．また，比較対象として個人ウェブサイト上の作品紹介文も掲載する．

・生成文章例 1

舞台重視，ネタバレ度 1(冒頭のみ)

主人公と同僚の白木はリスボンからジブラルタル沖のゼルシー島に向かい，メントール公爵の城塞に潜入する．

ゼルシー島は極楽島とも称される魅力あふれる島であるが，白木によると，メントール侯爵はある重要な機密を隠しているという．

国際的な舞台で展開するスパイ事件が主題である．

・生成文章例 2

人物重視，ネタバレ度 2(核心以外)

主人公は温和な性格ながら日本の密偵である．彼は勇猛かつ突拍子もない同僚白木と共にイギリス軍の機密に関わる任務に臨む．

機密の鍵を握るメントール公爵は不思議な音叉の力を用いる謎多い人物である．また，幾多の密偵を殺害してきたネッソンという牧師も主人公を阻む．

主人公達は無事に任務を達成することができるのか．

緊迫感とユーモアのある人物と会話が特徴である．

・比較対象文

(出典)http://www.geocities.jp/web_hon/01/unno.htm

『暗号音盤事件』（青空文庫）

短編。ジブラルタルから南西へ千キロにあるマデイラ諸島の小さな島・ゼルシー島へやって来た日本人の密偵二人。イギリス軍の重要機密である暗号の鍵を探し出すため、メントール侯の城塞へ潜入した二人は、敵が迫り来る中、目をつけた音盤(レコード)を一枚一枚調べるが…。「これも駄目か。が――待てよ」。国際スパイ小説。

3.2 課題:短期的課題

前述の通り，現在の研究は「テキスト解析によるデータベース構築」と「作品推薦文生成システム」を用いて進行し，システム作成の結果，利用者にとって必要な情報を適切に盛り込んだ作品推薦文の作成にはある程度成功したといえる．しかし，現状の方式，特にテキスト解析によるデータベース構築には 課題も複数挙げられる．

一つ目の課題は「作品データベースの量的・質的充実」である．既に作品取得を行いデータベース構築を進めた作品であっても，文体や内容の整理を済ませたものから単純な統計解析を済ませたに留まっている状態のものまでデータベースの充実度にばらつきがあるため，それらの充実度を揃えるとともに作品数自体も増強する必要がある．

次なる課題は「作品情報抽出の自動化・効率化の推進」である．現状，手動に大きく依拠している上に必要な作業量の多い作品情報の抽出部分について，より迅速な抽出方式や自動的な内容整理を行わなければ作品データベースの拡充も望めない．

また，現状の方式では「抽出内容の妥当性」についても問題が発生しうる．抽出を行った研究者個人の重視する評価基準と一般に重視される評価基準が一致するとは限らない．

現在検討中のこれらの課題解決の一つの方法として「感想データベースの活用」が考えられる．テキスト外情報や，個人の観点に偏らない多面的な視点からの分析が期待できる反面，採用段階での偏向や少数意見の脱落などには注意が必要となる．

3.3 課題:長期的課題

ここでは，作品推薦システム構築の先に存在すると考えられる課題について検討し解決方法について構想する．

「作品推薦システムの対人性能評価とフィードバック」に関しては，現在想定している学生を対象とした性能評価のみならず，多様な年齢層への対応，Web 上でのシステム効果による広範囲な対象からの対人性能評価と課題指摘を受け，それらをシステムへフィードバックし改良する必要があると考える．

以下に述べるのは更なる長期的課題，データベース構築と推薦文生成システムが実用域に達した段階で考慮する課題である．

「著作権課題」については，特に不特定テキストに対する応用の際に問題になる．実際に内容評価を必要とする文章は商業出版されている書籍が多い．推薦文は直接文章の内容を開示しないが，解析の段階で文章全体へのアクセスは可能でなければならない．どのように著作権保護項目へのアクセスを得るか，侵害を少なく留めるかは重要な課題である．感想データベースを活用する場合も，感想や評価文の利用権について考慮する必要があり，決してハードルは文章そのものに比べて低いとはいえない．

「複合メディアへの対応」も重要な課題である．特にインターネットメディアでは画像だけでなく音声,動画の付属するコンテンツも一般的である．付属メディアの視聴を前提とする文章も少なくなく，自動的な理解には画像（音声,動画）認識技術との複合も必要になると考えられる．

参考文献

[難波 2006] 難波英嗣: 情報抽出を利用した複数文書要約,知能と情報 vol.18, No.5, 日本知能情報ファジィ学会

[原田 2011] 原田隆史: 感性パラメータを用いた類似する小説の提示,情報知識学会誌 2011 vol.21, No.2

[松浦 2012] 松浦有容・渥美幸雄: 感情表現による書籍情報の可視化手法の提案と実装，専修大学情報学研究所所報 vol.78

[Vishal 2010] Vishal Gupta, Gurpreet Singh Lehal : A Survey of Text Summarization Extractive Techniques, Journal of Emerging Technologies in Web Intelligence. Vol.2, No.3

[Dipanjan 2007] Dipanjan Das, Andre F.T. Martins : A Suevey on Automatic Text Summarization, Language Technologies Institute, Carnegie Mellon University.

[泉谷 2009] 泉谷知範・前田栄作: アクティブ探索法の探索効率と探索精度に関する一考察，日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

[那須川 2001] 那須川哲哉・河野浩之・有村博紀 : テキストマイニング基盤技術，人工知能学会誌 16 巻 2 号.

[市村 2001] 市村由美・長谷川隆明・渡辺勇・佐藤光弘: テキストマイニング-事例紹介，人工知能学会誌 16 巻 2 号.