自由記述文書データからの知識発見手法に関する研究

A Study of Knowledge Discovery Method from Free Format Text Data

渡辺　智幸

WATANABE, Tomoyuki

概要：近年の情報技術の発展に伴い、自由記述アンケートや製品に対するユーザーレビューなどの文書データを大量に収集、蓄積することが容易となっている。しかし、人間が大量の文書データに目を通し、分類整理を行うことは多大な時間と労力を要するため、分析を自動化する方法論が求められている。本研究では、テキストマイニングの技法に基づき、大量の自由記述アンケート及びユーザーレビューから知識発見を行うための手法について検討する。分析対象とする文書データの種類や性質に応じて、それに有効な手法も異なると考えられることから、これら 2 種類の文書データに対し、異なる知識発見手法を提案する。分析実験を通じて、各提案手法の有効性を評価し、考察を与える。

Summary: Recently, it has been easy to collect and accumulate huge amount of free format text data by development of information technology. For example, free format questionnaire or review for users about products can be collected as the format of digital data by using information technology. However, it may be very hard to look over and analyze enormous text data by person. Therefore, methodology for automatic analysis is needed to get knowledge from huge amount of text data in practice. This paper proposes new methods to discover effective knowledge from enormous free format questionnaire or review for users based on text mining approach. In this paper, two different methods for knowledge discovery are proposed with respect to the purpose of analysis. Finally, effectiveness of proposed methods is evaluated by the application and analysis experiment.

キーワード: 自由記述文書, 情報検索, 自然言語処理, 知識構造図

Keywords: Free Format Document, Information Retrieval, Natural Language Processing, Knowledge Structure Diagram

１．研究背景

近年、情報技術の発展に伴う情報の電子化により、インターネットを通じて大量の自由記述文書データを収集、蓄積することが容易となっている。電子データという形で収集された自由記述アンケートやユーザーレビューには、企業などにとって非常に有益な情報が含まれている可能性が高い。従って、これらの情報を有効に活用していくことが求められるが、獲得できる文書データ量が膨大であるため、人手で全てを適切に処理することは多大な時間と労力を必要とする。

大量の文書データを有効に活用するには、少量の負担で必要な情報を効率よく引き出すための手段が必要であり、この要求を満たすものが情報検索や自然言語処理の技術である。これらは現在ではテキストマイニングの技法へと発展を遂げ、文書データからの知識発見や文書の自動要約といった手法が実用に近づいている。

２．研究目的

企業などが、顧客の自由意見を収集する場合に代表的な手段として、自由記述アンケートやユーザーレビューの収集がある。ただし、これらの文書データの分析を行う場合、対象とする文書データの種類や性質に応じて、それに有効な手法も異なると考えられる。本研究では、上述した自由記述アンケートとユーザーレビューの 2 種類の文書データを対象として、以下に示す異なる提案手法を用いた知識発見を行う。

① 自由記述アンケートからの知識発見

ベクトル空間モデルにおける文書間の類似度の概念に基づき、大量のアンケートデータを集約して代表的な意見のみを自動的に抽出する手法を提案する。

② ユーザーレビューからの知識発見

あらかじめ人間が作成した知識構造に対応した単語の出現頻度に基づいて知識発見を行う手法を提案する。人間が作成した知識構造とは、特定の分野に関する自由記述文書に出現する単語を意味の類似性によって分類し、各単語群の関係性を導き出して構造化したものを指す。

以上のように、①, ②の問題に適応した知識発見手法を提案し、分析実験を通じてその有効性を評価することを本研究の目的とする。

３．自由記述アンケートからの知識発見

３－１．概要

分析者が自由記述アンケートの自動分析結果に求める情報は、「どのような意見が代表的で、同類意見がどの程度存在するのか」という結果である。本節では、ベクトル空間モデルに基づく文書間の類似度計算によって文書の重要度を算出し、このような知識発見を可能とする手法を提案する。提案手法の適用範囲は、「～について自由

この研究の一部は、情報処理学会 FIT2005 第 4 回情報科学技術フォーラムと日本経営工学会の平成 18 年度春季大会，秋季大会、平成 19 年度春季大会，秋季大会において発表を行った。

にお書きください」といったある程度回答内容が類似しやすい自由記述アンケートデータである。

３－２．文書の類似度・重要度の算出法

本節で用いる基本的なベクトル空間モデルによる文書の類似度・重要度の算出法を以下に示す。

[定義１] 文書ベクトル

文書 x における各単語の出現回数を要素とするベクトルを文書ベクトル v_x とする。

[定義２] 文書の類似度

文書 x,y の類似度 sim(x,y)を単語ベクトル v_x,v_yの余弦で定義する。T はベクトルの転置を表す。

$$sim(x, y) = \frac{v_x \cdot v_y{}^T}{\sqrt{(v_x \cdot v_x{}^T)(v_y \cdot v_y{}^T)}} \qquad (1)$$

[定義３] 文書の重要度

亀田の手法[1]に従い、多くの文書と類似度の大きい文書は、文書集合全体の多くの内容を含んでいると考え、文書間の類似度の平均により文書の重要度を与える。

文書 x の重要度 imp(x)を次式で定義する。

$$imp(x) = \frac{1}{m-1}\sum_{y \neq x} sim(x, y) \qquad (2)$$

ただし、m は文書の総数である。

これら 3 つの定義を用いて、重要度 imp(x)が大きい順に文書を抽出する。

３－３．従来手法

従来研究[2]では、類似度閾値γというパラメータを用いて、重要文書が 1 つ抽出されるたびに、その文書との類似度がγ以上である文書を抽出候補リストから除外し、残った文書の中から次の重要文書を抽出する。これにより、内容の類似した文書が抽出されにくくなる。除外された文書は類似文書として扱い、その数を表示することで抽出された文書の代表性の指標とする。γの値はユーザーが与えるパラメータであり、$0 < \gamma \leqq 1$ とする。

【従来手法アルゴリズム】

Step1 分析対象となる文書集合を形態素解析して単語を抽出し、各文書を文書－単語ベクトルで表す。

Step2 全文書について、各文書間の類似度 sim(x,y)を式(1)より計算する。

Step3 全文書について、各文書の重要度 imp(x)を式(2)より計算する。

Step4 未抽出文書のうち重要度が最も大きい 1 文書を抽出する。

Step5 抽出した文書との類似度がγ以上の文書を抽出候補リストから除外し、類似文書として扱う。類似文書数を数えて抽出文書の付随情報とする。

Step6 残り全ての文書が除外された場合は終了。

Step7 Step4 に戻り、残った文書の中から重要度が最も大きい 1 文書を抽出する。 □

３－４．提案手法

抽出した重要文書と類似文書群の中で、類似度・重要度の再計算を行い、代表文書を改めて抽出するという処理を従来のアルゴリズムに追加する。これにより、抽出された重要文書とその類似文書が本当に類似しているか、文書全体または類似文書群の代表としてふさわしい文書かなどを検証する。

【提案手法アルゴリズム】

Step1 分析対象となる文書集合を形態素解析して単語を抽出し、各文書を文書－単語ベクトルで表す。

Step2 全文書について、各文書間の類似度 sim(x,y)を式(1)より計算する。

Step3 全文書について、各文書の重要度 imp(x)を式(2)より計算する。

Step4 未抽出文書のうち重要度が最も大きい 1 文書を抽出する。

Step5 抽出した文書との類似度がγ以上の文書を抽出候補リストから除外し、類似文書として扱う。類似文書数を数えて抽出文書の付随情報とする。

Step6 抽出された重要文書に対する類似文書群内で類似度・重要度を再計算し、代表文書を改めて抽出する。類似意見がなければ、元の文書をそのまま代表とする。

Step7 残り全ての文書が除外された場合は終了。

Step8 Step4 に戻り, 残った文書の中から再び重要度が最も高いものを抽出する。 □

３－５．自由記述アンケートの自動分析実験

３－５－１．使用アンケートデータ概要

分析実験に使用したアンケートの概要を以下に示す。

【実証実験後の Web 科目登録システムの性能と利用環境に関するアンケート】 設問： Web 科目登録システムを利用した感想をお聞かせください。(自由記述) 有効回答数：607 名

３－５－２．実験方法

従来手法と提案手法それぞれで重要文書ランキングの抽出を行い、各手法における抽出結果の上位 30 件の内容を比較する。そして、抽出結果の差異や情報多様性などを検証する。この時、類似度閾値γは抽出文書数が最小となる値を探索的に設定する。その値は分析対象データに応じて一意に決定されるものであり、この実験においては $\gamma = 0.0001$ となった。

３－５－３．結果の考察

従来手法と提案手法による重要文書抽出結果の一部を図 1、図 2 に示す。2 つの結果を比較すると、重要文書上位 30 件において差異が確認できる。2 段階の重要文書抽出を行う提案手法では、上位 30 件中 15 件が 2 回目において 1 回目とは異なる文書を抽出している。そしてそれ

らは、1 回目に抽出された内容を含んだ長めの文書であることが多い。つまり、提案手法は1回目に抽出した内容をある程度保持したまま、総合的に得られる情報量を増加させることに成功していると言える。また、それに関連して、従来手法では Web 科目登録システムの性能について客観的な事実を述べているだけの文書が多く抽出されていることが分かった。一方、提案手法では客観的事実に加えて主観的な感想やシステムに対する要望などが含まれる文書が多く抽出されているため、アンケート分析者にとってより有益な結果が得られたと言える。これは、提案手法の 2 段階重要文書抽出において、長い文書が抽出されやすい性質と、長い文書には貴重な情報が含まれている可能性が高いという事実がうまく噛み合った結果と考えられる。

順位	類似	原文
1	376	この科目登録は、便利で科目登録がスムーズに行えると思う。
2	71	科目登録中、特に問題は起こらず、スムーズに登録することができました。
3	27	とても簡単に科目の検索ができて、そのうえ登録もスムーズに行えたので使いやすかったです。
4	2	前回の実証実験に比べ極めて短時間で科目登録が行えた。これくらいスムーズに登録できるなら
5	35	今回は、全体としてはスムーズに科目が登録できたのでよかったと思います。
6	0	MEB実験には何度か参加しているのですが、今回はスムーズに進んだので、これだったら実際の
7	0	自宅から科目登録できるので、とても便利だと思う。
8	1	かなりスムーズに登録ができて、実際に使えたらかなり登録が楽になると思う。予想よりも、時
9	9	自分のとりたい科目が表示されないと困ったが、慣れると簡単に利用できて登録もスムーズにで
10	0	意外にスムーズにできたが、本当に登録されているかが不安。科目登録には期限があるためWeb
11	0	実際に登録をした時や前回の実験に比べ、格段にスムーズに行うことが出来た。これならばWeb
12	0	前回は再試行や接続できない状態が頻繁にでていたがそれは今回はなかったのでスムーズに科目
13	0	自宅で科目登録できると落ち着いてできるので非常にいいと思う。ただ再試行が出たり処理が
14	0	なかなか科目が登録できず、最後の方は少しいらいらしてしまいましたが、webでの登録によっ
15	5	１回もエラーが出ることなく、科目登録できた。
16	4	比較的すいている時間にアクセスできたので、スムーズに登録作業を行うことができました。
17	0	特に時間がかかるということもなく、自宅から登録できるので便利だと思います。
18	2	前回の実験時とは異なり、非常にスムーズでストレスも感じることなく、きちんと登録するこ
19	0	簡単でいいと思う。自分が選んだ科目が申請科目として一覧でてるから、登録ミスとかも減ると
20	0	わざわざ大学まで行かなくても科目登録でき，しかも夜なでも申請できるので，大変よい。

図 1. 従来手法の重要文書抽出結果（1 段階抽出）

順位	類似	原文
1	376	スムーズに科目登録できたのでよかった。
2	71	思った以上にスムーズに科目登録をすることができた。2005年度の科目登録でWebシステム
3	27	スムーズに行えた。実際の科目登録でも、Ｗｅｂシステムを利用できたら、自宅からも操
4	2	今回の実験で行ったように登録できるのであれば、十分に使えると思います。
5	35	簡単に出来て、時間もあまりかからずにしかも自宅で科目登録が出来てすごく便利でした。
6	0	MEB実験には何度か参加しているのですが、今回はスムーズに進んだので、これだったら実
7	0	自宅から科目登録できるので、とても便利だと思う。
8	1	思っていたよりスムーズに登録出来たので良かった。
9	9	科目の区分、例えば関連科目とかコア科目などはいちいち講義要綱を見ないとどの科目が
10	0	意外にスムーズにできたが、本当に登録されているかが不安。科目登録には期限があるた
11	0	実際に登録をした時や前回の実験に比べ、格段にスムーズに行うことが出来た。これなら
12	0	前回は再試行や接続できない状態が頻繁にでていたがそれは今回はなかったのでスムーズ
13	0	自宅で科目登録できると落ち着いてできるので非常にいいと思う。ただ再試行が出たり処
14	0	なかなか科目が登録できず、最後の方は少しいらいらしてしまいましたが、webでの登録に
15	5	１回もエラーが出ることなく、科目登録できた。
16	4	これまで科目登録Webシステムを利用してきた中で今回の実験が最もスムーズに科目登録を
17	0	以前の実験のときよりもスムーズにできよかった。しかし、まだ不安が残るので引き続き
18	2	前回の実験時とは異なり、非常にスムーズでストレスも感じることなく、きちんと登録す
19	0	今回は前回の反省を活かせてスムーズな登録が行えた。同じ科目を選択したときに警告が
20	0	わざわざ大学まで行かなくても科目登録でき，しかも夜なでも申請できるので，大変よ

図 2. 提案手法の重要文書抽出結果（2 段階抽出）

４．ユーザーレビューからの知識発見

４－１．概要

自由記述文書データの自動分析手法は、今日ではマーケティングやブランドマネジメントなどの特定の目的に特化した利用も期待されている。従来研究[3]では、テキストマイニング技術を用いた顧客ロイヤルティの構造分析が行われている。また、何らかの目的に特化した自由記述文書データの分析を行う場合、従来の単語出現頻度を用いた手法に人間が作成した知識構造を援用することが有効である[4]。しかしながら、与えられた知識構造を用いてユーザーレビューの自動分析を行う場合には、より専門家による分析に近い、すなわち精度の高い方法を検討する余地がある。

本節では、ユーザーレビューの自動分析を支援するための、人間の判断に基づく肯定・否定の概念を取り入れた新たな知識構造化手法を提案する。前述の顧客ロイヤルティ構造分析を題材に、ユーザーレビューの自動分析を行い、人手による分析結果に近い知識発見が可能となることを示す。

４－２．従来の知識構造化手法

ここでは､従来研究[3]で示されている知識構造化手法について述べる。使用するデータは、化粧品専用のクチコミサイトである「@cosme」(http://www.cosme.net/)に顧客から寄せられた大量のユーザーレビューである。「@cosme」のユーザーレビューは､満足度を 7 段階で評価し、製品についての感想を自由に記述する形式となっている。その中から、「基礎化粧品」、「ベースメイク」、「メイクアップ」の 3 カテゴリより各 150 件ずつ、合計 450 件を抽出している。ユーザーレビューの知識構造化によって作成した顧客ロイヤルティ構造図を用いた自動分析を行うことで，化粧品購買者の顧客ロイヤルティ構造を明らかにするという狙いである。構造図の作成手順を以下に示す。

【従来研究における知識構造図の作成手順】

Step1　クチコミ評判サイトからユーザーレビューを収集する。

Step2　抽出したユーザーレビューを形態素解析により単語に分割する。

Step3　分割した単語から不要語を除去する。

Step4　意味が類似している単語同士を KJ 法で分類する。

Step5　分類した単語群を構造化する。　□

上記の手順で作成された化粧品購買者の顧客ロイヤルティ構造図が次の表 1 である。この顧客ロイヤルティ構造図は、大中小の各項目から成り立っており、それぞれがロイヤルティの構成要素を示している。

表 1. 化粧品購買者の顧客ロイヤルティ構造図

	大項目	中項目		小項目
顧客ロイヤルティ	便益	品質全般	機能性	効果
				成分
				色
				持続性
				即効性
				使用外見
			外的特性	パッケージ
				内容量
			独自性	ブランド価値
				限定
		個人特性	意思決定	購入意思
				比較
				評判
			価値観	必要性
				相性
				推薦
			感情	好み
				使用感
			使用用途	使用方法
				ライフスタイル
			利便性	便利さ
			習慣性	従来からの使用
			不確実性	初めての経験
	コスト	費用		価格
				割引
				コストパフォーマンス
		流通		購入方法
	信頼性	信頼性向上		試用
				メディアの影響
				身近な人の影響

４－３．肯定・否定の概念を用いた知識構造の作成手法

4-2 で示した構造図は、形態素解析によってあらかじめ分割された単語に基づいて作成されている。そのため，肯定・否定などの文脈を考慮することができないというデメリットが存在する。これを用いてユーザーレビューの自動分析を行った場合、単語の出現頻度から多くの顧客が言及している構造図の要素はどれかという情報は得られるが、それが肯定的な内容か否定的な内容かまでは判断できない。しかしながら、人間の書く文章を考えた場合、同じ内容に言及していても、それが肯定的か否定的かによって、製品に対する評価の意味合いは全く異なる。このような制約のため、従来手法で作成された表１のような構造図を用いたユーザーレビューの自動分析は、しばしば人間による判断結果と異なってしまっていた。すなわち、構造図も肯定・否定に分類されていた方が結果の活用の際に好ましいと考えられる。本節では、人間の判断に基づく肯定・否定の概念を取り入れた知識構造の作成手法を提案する。

ここでは、肯定（Positive)、否定（Negative）による分類を基本に、どちらとも取れる単語については共通（Common)とする CPN 分類を用いた知識構造を作成した。その作成手順を以下に示す。

【CPN 分類を用いた知識構造の作成手順】

Step1 収集したユーザーレビューを分析者が読み、構造図の各小項目について肯定的な内容の記述、否定的な内容の記述、それ以外の記述に分類する。

Step2 各小項目に分類された文章を形態素解析により単語に分割する。

Step3 分割した単語から不要語を除去する。

Step4 同小項目内で肯定・否定の両方に出現する単語を共通単語として分類する。 □

上記の手順で作成した CPN 単語分類リストの一部が次の表 2 である。この CPN 分類を用いた知識構造を利用してユーザーレビューの自動分析を行うことで、顧客が製品のどの要素に関して言及しているかだけでなく、それが肯定的か否定的かという内容まで把握することができると考えられる。この結果は、ユーザーレビューの内容を企業が製品戦略などに活用する際、大きな手助けとなることが期待される。

表 2. 単語の CPN 分類例

効果			成分		
C	P	N	C	P	N
UV	GOOD	あんまり	SPF	オリーブ	ザラザラ
WP	あっぱれ	いまいち	オイル	完璧	悪い
カバー	いい	かゆい	テクスチャ	重要	難点
キープ	うるおう	くすむ	安全	無臭	肌荒れ
落ちる	すごい	しみる	添加	良好	負担

C＝共通　P＝肯定　N＝否定

４－４．CPN 分類を用いた知識構造図に基づくユーザーレビューの自動分析実験

各ユーザーレビューについて、構造図の小項目 CPN に当てはまる単語が存在すれば 1、存在しなければ 0 を立てるというアルゴリズムに基づき、コンピュータで自動的に 0-1 行列を作成する（自動分析)。その一方で、人間がユーザーレビューを読んで理解し、手動で 0-1 行列を作成する（手動分析)。手動で作成した行列を正解データとして、自動で作成した行列との比較を行い、提案手法の精度を評価する。評価指標は正解率、再現率、適合率である。実験結果を表 3 に示す。

表 3. 自動分析の精度評価

	A	B	C	総合
正解率(%)	74.2	75.3	75.7	75.1
再現率(%)	84.7	70.9	78.2	77.8
適合率(%)	15.8	15.4	16.4	15.9

A：基礎化粧品　B：ベースメイク　C：メイクアップ

４－５．結果の考察

自動分析実験の結果、手動分析に対して約 75%の正解率が得られた。ユーザーレビューで顧客が言及している内容と、その肯定・否定を把握するという目的に対してある程度の有効性が示された。問題点としては適合率の低さが挙げられる。提案した知識構造は、複数の小項目に同一の単語が存在することが多く、自動分析時に複数の項目に１が立つ現象が起きやすい。つまり、正解データにて頻繁に出現する要素については高い再現性が期待できるが、それ以外の要素はやや誇張された結果となる傾向がある。手動分析の精度に近づけるためには、知識構造に登録する単語を厳選して適合率を上昇させ、再現率とのバランスを調整していくことが必要と考えられる。

５．まとめ

本研究では、自由記述アンケートとユーザーレビューを対象に、文書データの種類や性質に応じた異なる知識発見手法を提案し、自動分析実験を通じて評価を行った。その結果、分析者がそれぞれの自由記述文書データから獲得したい知識を発見でき、かつ従来に比べ人手による分析に近い結果が得られることが示された。

参考文献

[1] 亀田雅之：“擬似キーワードによる重要キーワードと重要文の抽出”，言語処理学会第 2 回年次大会発表論文集，pp.97-100, (1996)

[2] 渡辺智幸, 後藤正幸, 石田崇, 平澤茂一：“情報検索技術を用いたアンケートデータの分析手法に関する研究”，日本経営工学会平成 18 年度春季大会予稿集, pp.126-127, (2006)

[3] 三川健太, 高橋勉, 後藤正幸：“テキストデータに基づく顧客ロイヤルティの構造分析手法に関する一考察”, 日本経営工学会論文誌, Vol. 58, No. 3, pp.183-192, (2007)

[4] 渡辺智幸, 三川健太, 海老澤卓哉, 後藤正幸：“知識構造を利用した文書データの自動分析に関する一考察”, 日本経営工学会平成 19 年度春季大会予稿集, pp.48-49, (2007)