ウェブ型テキストマイニングツールiTMの開発

誌名	システム農学
ISSN	09137548
著者名	朴,壽永 長谷部,正 安江,紘幸
発行元	システム農学会
巻/号	32巻1号
巻号補足	
掲載ページ	p. 25-35
発行年月	2016年1月

技術論文

ウェブ型テキストマイニングツール iTM の開発

朴 壽永 [1)]・長谷部 正 [1)]・安江紘幸 [2)]

1)東京農業大学
2)農研機構東北農業研究センター

(受付日 2015/04/07; 受理日 2015/12/23)

要旨

近年、日本の農業部門においてもテキストマイニングによる研究の試みがあり、その対象も多岐にわたるが、学術文献の数はまだ少ない。その理由として、テキストの形態素解析のために情報システムが必要であること、抽出された形態素を統計分析に用いるための数値化が大がかりの作業であることがあげられる。そこで本稿では、研究者のみならず担い手農業者などのインターネットユーザが普段慣れているウェブサイト上のボタンをクリックするだけで日本語形態素解析に加え、構文解析やカイ2乗検定・フィッシャーの正確確率検定などを順次行うことができるウェブ型テキストマイニングツールの iTM(internet Text Mining)を開発した。システムの実用性を確かめるために、iTM に備え付けのウェブ型アンケート調査ツールを活用、取得した自由記述回答文を対象に事例分析を行った。その結果、iTM の特徴であるウェブブラウザ上のボタンクリック機能によって、形態素解析エンジンなどのインストールや煩雑なデータの数値化作業がなくなり、気軽にテキストマイニングを実施できることが明らかになった。なお、繰り返し出現する単語の評価法として独自で考案した $1 \leqq Y < 2$ 評価法を適用・比較することで、分析結果に対する考察をさらに深められることが示唆された。

キーワード

形態素解析、構文解析、差の検定、出現度数評価法、農業

1. はじめに

情報システムの発展と普及によってテキストデータが急速に増え続けている。それに伴うテキストマイニングの必要性も増していると言える。また、選択肢によるアンケート調査は調査者の意図が反映されやすく、情報探索的な調査や分析が困難であるため、自由記述による設問項目を設ける場合が多い。しかし、せっかく得られた自由記述の回答文は分析されることなく、そのまま放置されるケースも少なくないのが現状である。

こうした中で、近年、日本の農業部門においても消費者ニーズの把握(磯島 2006)を端緒として、企業の農地利用(堀部 2014)や地域づくり(木南ほか 2013)、農村移住者の農村像(飯坂 2013)などのテキストマイニングによる研究の試みがあり、その対象も多様であるが、文献の数はまだ多くない[注 1]。テキストマイニング手法による研究が少ない理由として、テキストの形態素解析[注 2]や構文解析[注 3]のための情報システムが必要であること、抽出された形態素を統計分析に用いるための数値化が大がかりの作業であることがあげられる。いくつかの商用ソフトがあるが、汎用ソフトよりはるかに高額である(IT トレンド 2015) 。また、詳細は表 2 に後述するが、汎用ソフトは形態素解析エンジンなどのインストールやインターフェース・入力作業が煩雑であるため、活用が簡単ではないのが現状である。ウェブ型アンケート調査代行と分析ツールを提供する有名大手リサーチ専門会社もテキストマイニングに関しては、回答者属性にリンクさせた回答文の羅列を提示するだけに留まっており、テキストマイニングの難しさが伺える。

以上のことから、本稿では研究者のみならず担い手農業者も簡単にテキストマイニングができることを目的としたウェブ型テキストマイニングツールの iTM(internet

1) 〒156-8502 東京都世田谷区桜丘 1-1
2) 〒020-0198 岩手県盛岡市下厨川字赤平 4
(Correspondence: s3paku@nodai.ac.jp)

Text Mining)を開発し、事例分析に適用することでその実用性を確かめた。本稿で紹介するシステムは http://www.bumoc.net サイト上から無料で活用できる。以下、2節にiTMの概要と特徴を述べ、3節では機能の詳細を整理した。4節には iTM の活用による事例分析を2つ示した。1つ目は iTM の備え付け機能のみで行った事例、2つ目は iTM から出力した数値データを別途統計分析ソフトの SPSS に適用して行った事例である。

2. iTM の概要と2つの特徴

2.1 iTM の概要

プログラミング言語は PHP、JavaScript、JQuery、HTML5、CSS3を、データベースはMySQLを用いた(表1)。オープンソースの日本語形態素解析エンジンであるMeCab(2015)を独自のサーバにインストールし、出された課題や意見の文字列を対象に品詞を解析した。助詞と副詞を除き、名詞・形容詞・動詞・助動詞の文字や文字列頻出傾向の度合を解析した。構文解析は Yahoo Japan (2015) の日本語係り受け解析 API (Application Program-ming Interface) [注4]を活用した。グラフの作成は Google Charts(2015)の API を活用した。

2.2 iTM の2つの特徴

2.2.1 ウェブブラウザ上のボタンクリックで実行できる機能

図1にiTMの処理フローを示した。すべての機能をマウス操作で利用でき、例えば、マウスで属性選択後1回、ウェブブラウザ上でボタンをクリックすると形態要素解析の結果が、また、分析対象のキーワード選択後1回、ボタンをクリックすると構文解析の結果が得られる。

図2に磯島(2009)の処理フローを示した。図2の点線枠内の処理内容は図 1 の「形態素解析」と「選択キーワードを含んだ文章の閲覧」に当たるもので、iTM では 2 回のボタンクリックだけで同様の処理結果が得られる。磯島によるプロセスでは、Excel などを用いた煩雑かつ高度な手作業が必要とされるため、人為的ミスもありうる。他にも、表 2 のように、統計分析機能が豊富であることから研究に活用されている KH Coder(2015)などがあるが、KH Coder は、構文解析機能がなく、高機能であるために誰でも簡単に使いこなせる UI (User Interface)とは言

表1 開発環境

サーバOS	Linux
Web サーバ	Apache
DBMS	MySQL 5.1.69
日本語形態素解析エンジン	php-mecab
サーバサイド言語	PHP Version 5.3.3
クライアントサイド言語	JavaScript、JQuery、HTML5、CSS3
動作確認のウェブブラウザ	Chrome 45.0.2454.101 m Internet Explorer 10 Firefox 4.0.1.167 Safari 6.2 iOS/AndroidOS (標準ブラウザ)

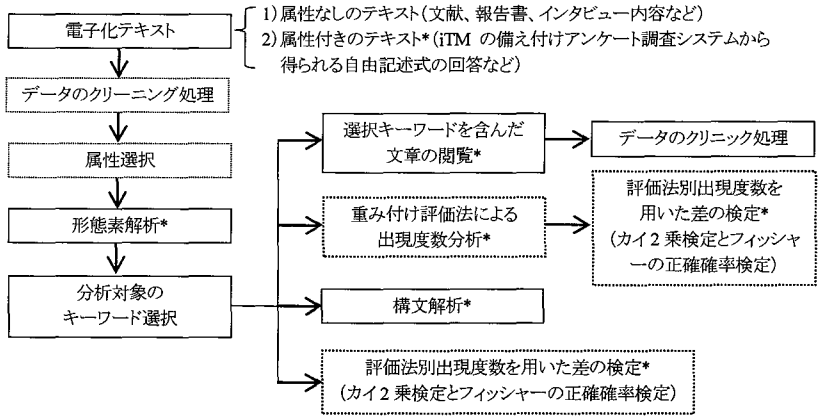

図1 iTM の処理フロー

[a]点線枠は、属性付きのテキストの場合のみ対応する機能である。
[b]*は、それぞれの処理結果を統計分析用 Excel ファイルでダウンロードできることを示す。

い難い。RMeCab（石田 2008）は、ユーザ側による R のコマンド語入力が必要とされる。石田（2008）によると、構文解析システムの RCaboCha の開発が試みられているが、完成の報告はまだない。形態素解析と構文解析の機能を有する TinyTextMiner（TMM）（松村ほか 2009）は統計分析の機能が乏しい。なお、これらの全てはソフトのインストールも必要であることなどから担い手農業者などが広く利用するにはまだ敷居が高いと言える。また、本研究のようなウェブ型システムはない。

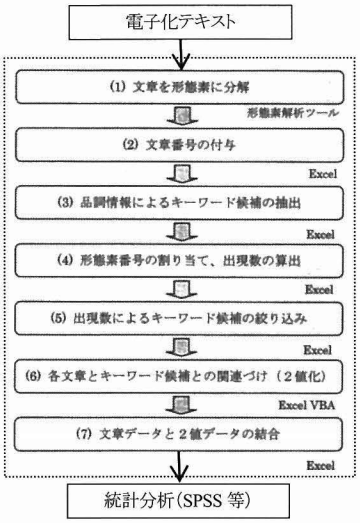

図 2　磯島（2009）の処理フロー

一方、iTM は、インターネットユーザが普段慣れているウェブサイト上のボタンをクリックすることで日本語形態素解析に加え、カイ 2 乗検定・フィッシャーの正確確率検定や構文解析の結果などが順次得られるため、後述の 4 節 4.1 のように iTM の活用のみでも簡単にテキストマイニングを実施できる。また、タブレットやスマートフォンからも利用できるとともに、独自の統計分析が必要な場合には統計解析ソフトでの分析が可能になるよう数値データの出力機能も有する。

iTM がウェブ型システムであることの意義は次の理由で非常に大きいと言える。ウェブ型システムは、UI を含めたアプリケーションの機能がすべてサーバ側で実現でき、ユーザ側にはウェブブラウザだけがあればよい。そして、一般ユーザが最も慣れている UI はウェブブラウザであるからである。これらは、これまで敷居が高かったテキストマイニングを、iTM を用いることによって、ユーザがウェブブラウザ上のボタンクリックでテキストマイニングを気軽に実現できることを意味する。

2.2.2　形態素の出現度数評価法

一人の参加者により出された意見には単語が繰り返して現れることもあり、出現度数がそのまま利用されると、その単語のもつ重要度が異なることがありうる。このことから既存の評価法としては、繰り返し単語をそのままカウ

表 2　既往システムと iTM の比較

システム名	システムの種類	インストール	利用媒体	備え付け機能				対応言語	使いやすさ	開発年度
				形態素解析	構文分析	属性別分析	統計分析			
KH Coder	スタンドアローン型	KH Coder	パソコン	○	×	△	○	日本語英語	△	2004
磯島の Excel 利用型	スタンドアローン型	Chasen	パソコン	○	×	×	×	日本語	×	2006
RMeCab	スタンドアローン型	RMeCab R、MeCab	パソコン	○	×	△	×	日本語	×	2008
TTM	スタンドアローン型	TTM MeCab、Cabocha	パソコン	○	○	×	×	日本語	△	2009
iTM	ウェブ型	無	パソコン、タブレット、スマートフォン	○	○	○	△	日本語	○	-

[a]統計分析において、KH Coder は対応分析（数量化 III 類）・クラスター分析・多次元尺度構成法・自己組織化マップ・共起ネットワーク・機械学習（ナイーブベイズ）などに、iTM はカイ 2 乗検定、フィッシャーの正確確率検定に対応している。

[b]使いやすさは、筆者の主観的な評価によるものである。

ントする方法(以下、Y=n と記す)に加え、1 にカウントする方法(以下、Y=1 と記す)がある(三浦 2012, 磯島 2006)。すなわち、「1」か「全てカウント」というような両極的な評価法であり、本稿の 4 節 4.2 の事例分析の結果のように、考察結果が異なる危険性がある。そこで、iTM においては、次のように重み付け評価法を考案し、Y=n 及び Y=1 の評価法との比較分析を行った。

$$Y=2-(1/2)^{(n-1)} \quad (1)$$

ここで、n は出現度数、Y は重み付け後の出現度数を示す。Y は 2 を超えない(以下、$1 \leqq Y<2$ と記す)。

3. 機能の詳細

3.1 データのクリーニング機能

テキストの中の必要がない記号や誤った文字列を訂正したり、表現を統一したりするデータのクリーニングは、単純ではあるが非常に重要である(金 2009)。iTM では後述の図 5 の形態素解析結果から得られた単語を任意に選択しボタンをクリックすることで、図 3 のように当該単語を含む文章がすべてリストアップされ、ユーザは必要に応じて原文を参照しながら変更できる。

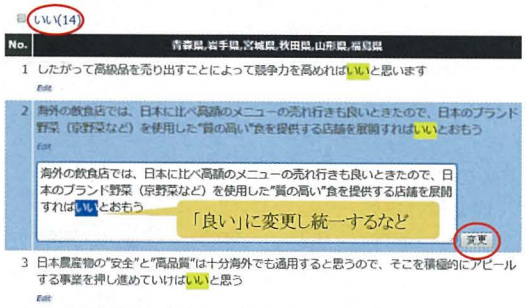

図 3 データのクリーニング

3.2 属性のグルーピング機能

属性付きのテキストについては、回答者数付きの属性ラベルが図 4 の左側のように表示される。属性別グルーピングは、任意の属性ラベルを右の「第1グループ」「第2グループ」にドラッグアンドドロップでできるため、感覚的なマウス操作でグルーピングを行える。属性別グルーピングは 2 つまでであるが、3 つ以上の属性別グルーピングは、後述の 4 節 4.2.2 のように第 1 のグループと第 2 のグループ、また第 2 と第 3、第 3 と第 1 のグループのように組み替えて行う。なお、2 つのグループにおける共通属性も指定できる。例えば、図 4 で「性別:男(85)」を共通属性に選択すると、男に絞ってグループ間の比較分析ができる。

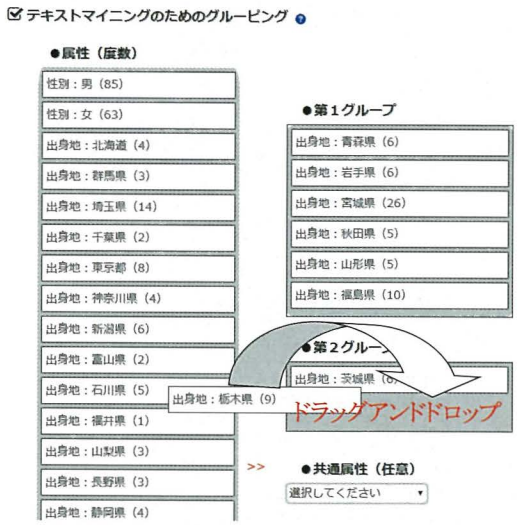

図 4 分析対象となる属性のグルーピング

[a]括弧内の数値は回答者数である。

3.3 形態素解析機能

図 5 に形態素解析結果を示した。図 4 でグルーピング

図 5 形態素解析

[a]数値は、2 つグループの出現度数の合計[第 1 グループの出現度数:第 2 グループの出現度数]を示す(Y=n)。

[b]紙面の都合上、中間部位のデータを切り取りし、画像編集を行った。

[c]48 位の単語は実名のため画像編集した。

後、実行ボタンをクリックすると図5が表示される。出現度数が 1 回以上ある単語に対して名詞のほか形容詞、動詞、助動詞がテーブルごとに整理される。

システムの処理速度を示すために、図4における実行ボタンをクリックしてからiTMのメイン画面である図5の形態素解析処理及び画面表示終了までのターンアラウンドタイム[注5]を計測した。その結果、4節4.1に用いた総文字数554のテキストは0.01227秒、4節4.2に用いた総文字数18,410のテキストは1.05651秒であった。計測時のWi-Fiインターネット回線の下り速度は7.8Mbpsであった。

3.4 構文解析機能

任意の単語（例えば、文化）をクリックすると当該単語を含んだ句を対象にした構文解析結果が得られ（図6）、当該単語を含む文節に対する前後の係り受け関係を把握できる。当該単語には赤字又は色が塗りつぶされているため内容把握がしやすい。また、図11と図12のように、係り受け関係の句に対する形態素解析結果が得られ、次のような考察を深めることができる（紙面の都合上、文節に対する形態素解析結果は省略）。例えば、「味が良いと思う」と「味が良いと思わない」のように、形態素解析を行うのみでは「何を良い」と評価しているかを知ることができず、また文脈によってニュアンスが異なる可能性を検証できない。つまり、ある語と係る語を一緒に取り出してその結びつきを明らかにすることが構文解析であり、その代表的な手法が係り受け解析である（松村ほか2009）。

☑ 「文化」を含む句の構文解析結果

文節	受け係り関係の文節（青森県,岩手県,宮城県,秋田県,山形県,福島県）
・和食が	-> 登録されたので、
ユネスコの	-> 無形文化遺産に
無形文化遺産に	-> 登録されたので、
登録されたので、	-> もらう
もっと	-> 展開し、
和食事業を	-> 展開し、
展開し、	-> もらう
日本の	-> 農産物に
農産物に	-> 持って
興味を	-> 持って
持って	-> もらう
もらう	-> 文末
日本の	-> 農産物は、

図6　構文解析

[a]点線で囲んだのは、句点で区切られた当該単語を含む最短文章。

[b]実線の中の「ユネスコの」が「無形文化遺産に」に対する係り文節で、「登録されたので、」が「無形文化遺産に」に対する受け文節である。

一方、iTMの構文解析はYahoo JapanのAPIを活用しており、iTMとYahoo Japanの間における当該テキストの送受信は暗号化して行われるが、送信テキストは Yahoo Japan に蓄積され、統計分析など何らかの形での再利用や漏えいの可能性が否定できない。

3.5 グループと単語別出現度数に対する差の検定機能

図5において任意の単語に☑を入れ、左下のボタンをクリックすることでカイ2乗検定・フィッシャーの正確確率検定結果が得られる（図7と図9）。最大2千個の単語について同時にカイ2乗検定・フィッシャーの正確確率検定が実施できる。iTMではY=nとY=1、1≦Y<2の出現度数評価法に基づいた最大6千個のカイ2乗検定・フィッシャーの正確確率検定結果がボタンクリックで同時に取得できる。

→ 重み付けなし(Y=n)
・食材

キーワード	食材				
1	合計	○	×	X^2値	7.507
青森県,岩手県,宮城県,秋田県,山形県,福島県	58	4	54	X^2値出現確率	0.006
茨城県,栃木県,群馬県,埼玉県,千葉県,東京都,神奈川県	54	14	40	1%水準で有意差あり ***	

図7　グループと出現度数を用いたカイ2乗検定

3.6 Y=nとY=1、1≦Y<2評価法による出現度数の算出と比較機能

Y=nとY=1、1≦Y<2の出現度数は、図5のボタンをクリックすることで結果が得られる（紙面の都合上、図を省略）。

3.7 統計分析用Excelファイルの生成機能

以上の 3.1、3.3、3.4、3.5、3.6 のボタンクリックで得られる実行結果は当該画面上に Excel ファイルのダウンロードボタンが設けられており、ボタンクリックでダウンロードができ、独自の分析などで活用できる。

3.8 備え付けのアンケート調査システムの活用

なお、紙面の都合上、iTM に備え付けのウェブ型アンケート調査システムの詳細については省くが、自由記述式型のほか、属性別、3段階尺度、4段階尺度、5段階尺度、2者択型、3つ以上の項目の中で1つ選択型、複数選択型設問項目に対応できる。

4. 分析事例

4.1 iTMの機能のみ用いた分析事例

担い手農業者などがテキストマイニングを気軽に実践

できるようになるためには、図5と図6、図7などにおけるiTM のボタンクリック機能を活用するだけで一定の分析結果が得られることが望まれる。そこで、iTMの機能のみ利用した分析を行い、その有効性を試みた。

4.1.1 テキストデータの概要

事例分析に用いた表 3 のデータは、手作業でも検証できるようサンプル数20のシンプルなデータを筆者が任意で作成したものである。新商品の食嗜好に関する設問の Q2 と Q3 に対し、女は否定的に、A11 以外の男は肯定的に評価している。Q2 に対する差の検定は、iTM の備え付けアンケート調査システムの t 検定機能を活用すると、ボタンクリックだけで図 8 の結果が得られ、Q3 の概要を簡単につかむことができる。

4.1.2 差の検定と構文解析による男女の違い

次のような手順でQ3のテキストにおける男女の違いをiTMの機能のみで分析した。

①図 4 で男女にグルーピングし、Q3 のテキスト(総文字数554)と図5を活用して名詞30個、形容詞18個、動詞24個、助動詞3個、合計75個の形態素を得た。

②75個の形態素を対象にしたカイ2乗検定・フィッシ

No	アイディア項目					
	グループA	回答数	合計値	平均値	分散	F値
	グループB	回答数	合計値	平均値	分散	F値出現確率
	等分散判定	t値	t値出現確率	差の検定結果		
1	試食した○○の味についてあなたの全般的な評価は？					
	男	10	42	4.20	0.400	1.139
	女	10	23	2.30	0.456	0.425
	等分散	6.496	0.000	1%水準で有意差あり ***		

図 8 属性間の t 検定結果

表 3 構文解析用サンプルデータの設問内容と回答内容

設問内容			回答形式
Q1	性別		選択式
Q2	試食した○○の味についてあなたの全般的な評価は？		5段階尺度による選択式
Q3	試食した○○の味について自由に書いてください。		自由記述式
回答	Q1	Q2	Q3
A1	女	3	正直にいえば、ちょっとまずかったかな…
A2	女	3	新商品になるような新鮮さを感じ取れなかった。少し柔らかくした方がいいと思う。
A3	女	2	良い味とは思えない。新商品としては評価を得られないのではと思う。
A4	女	2	まずかったし、見た目もいいと思わない。残念ながら、私にとっては納得いかないものだった。
A5	女	2	おいしくなかった。デザインはいいと思ったが、味はぜんぜんよくない。
A6	女	1	いい味だとは思えなかった。
A7	女	3	美味しいとは思うけど、買い求めて食べようとは思えない味。
A8	女	3	納得できない微妙な味だと思った。どちらかというとまずかった。
A9	女	2	新商品としてはいまいち良い味とは思えない。厳しい表現かも・・・
A10	女	2	私にとっておいしいとは思わない。何とも言えないような味だった。
A11	男	3	微妙な味だった。おいしいとは思わないけど、まずいとも思わない。
A12	男	4	味がよい。デザインもよさそう。人にお薦めできる商品な気がする。
A13	男	5	今まで食べたことのない味で、新鮮だった。ぜひ商品化してほしいと思う。
A14	男	4	おいしく食べられた。また食べてみたいと思うほど良い味だった。
A15	男	4	大変おいしかった。いい味に仕上がったと思う。
A16	男	5	味が非常に良い。消費者が満足しそうな味だと思う。
A17	男	4	子供の頃食べていたお袋の味だったので、とても懐かしく感じた。
A18	男	4	おいしかった。見た目もよいと思う。
A19	男	5	非常に美味しいと思った。歯ごたえもよかった。
A20	男	4	まろやかで非常に美味しかった。

[a]5段階尺度は、5=非常にそう思う、4=そう思う、3=どちらともいえない、2=そう思わない、1=全くそう思わない、である。
[b]A13の網に囲んだ「ない」は動詞であり、下線の「ない」は助動詞である。

ャーの正確確率検定結果から、繰り返し単語を評価するために 1≦Y<2 を採用(図 9)し、男女間に有意差が認められた単語を 2 個抽出した。動詞:思え(0:4)、助動詞:ない(1.5:8.5)であった(括弧内は男:女の出現度数。小数点以下は、差の検定に際して四捨五入される)。

③有意差が認められた助動詞「ない」を図 5 で選択、ボタンをクリックすることで、図 10 と図 11 の構文解析結果を得た。図 10 をみると、男は、(おいしい・まずい)～(思わ)～(ない)、女は、(良い味・新商品としての評価・見た目もいい)～(思え・得られ・思わ)～(ない)である。図 11 の女において、文章内に「味」と「ない」の比率が最も高く、図 10 と構文解析の数値データである図 12 に目を通すと、見た目に対する評価などもあるが、主に味に関する女の否定的な評価として「ない」が用いられていることが確認された。「思え」は、同じく図 10 と図 11 で、3 つの(思え)～(ない)、1 つの(思え)～(なかっ)が確認でき、女による味に関する否定的な評価に用いられていた。

以上のことから、①から③までのボタンクリックだけで、形態素解析及び属性別の形態素を用いた差の検定結果が得られ、有意差のある形態素に焦点を当てた構文解析を行い、係り受け関係の句・文節、図を吟味することによって、当該形態素が持つ意味やニュアンスの結びつきを定量的に把握できることが示された。

4.2　1≦Y<2 評価法の有効性に関する分析事例

実践で行ったアンケート調査結果を用いて、1≦Y<2 評価法の有効性を調べることを目的とした分析を行った。本分析においては、iTM から出力した数値データを別途統計分析ソフトの SPSS に適用して行った。

4.2.1　テキストデータの概要

大学 1 年生を対象に、海外で業績を伸ばしている大手外食企業 G 社の T 社長を招き、講演を行った後、表 4 のアンケートを実施した。講演終了後に 10 分間の回答時間を設け、iTM に備え付けのウェブ型アンケート調査システムを用いて実施した。スマートフォンなどインターネットに接続できる媒体がない場合は、紙ベースで回答してもらい後で手入力した。回答者数は 150 人、そのうち紙ベースでの回答は 43 人であった。

4.2.2　コレスポンデンス分析による 3 つの出身地の違い

図 5 を活用し、Q3 における無回答者 3 人を除いた 147 人の回答文(総文字数 18,410)から、名詞 687 個、形容詞 62 個、動詞 360 個、助動詞 13 個、合計 1,112 個の形態素を得た。回答者の所属大学が仙台に位置している

→ 重み付け(1= <Y <2)

・ない

キーワード	ない				
1	合計	○	×	X^2値	7.425
男	11	2	9	X^2値出現確率	0.006
女	12	9	3	1%水準で有意差あり ***	

・なかっ

キーワード	なかっ				
2	合計	○	×	X^2値	3.529
男	10	0	10	X^2値出現確率	0.060
女	10	3	7	10%水準で有意差あり *	
各セル度数の中に0に近い値があるため、X^2検定には向いてなく、Fisherの正確確率検定が望ましいです。				Fisherの正確確率	0.211
				有意差無	

図 9　グループと出現度数を用いたカイ 2 乗検定とフィッシャーの正確確率検定

☑ 「ない」を含む句の構文解析結果

文節	受け係り関係の文節（男）
おいしいとは	-> 思わないけど、
思わないけど、	-> 思わない
まずいとも	-> 思わない
思わない	-> 文末

文節	受け係り関係の文節（女）
良い	-> 味とは
味とは	-> 思えない
思えない	-> 文末
新商品としては	-> 得られないのではと
評価を	-> 得られないのではと
得られないのではと	-> 思う
思う	-> 文末
まずかったし、	-> いいと
見た目も	-> いいと
いいと	-> 思わない

図 10　「ない」を含む句に対する構文解析結果

[a] 紙面の都合上、下段部の表示画面を省略した。

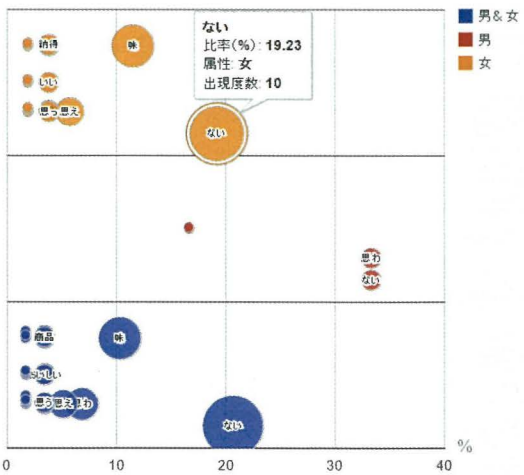

図 11　「ない」を含む句における出現単語と比率

[a]横軸は出現比率、縦軸は属性分けを示す。

[b]図中の出現度数は Y=n によるものである。

[c]図中の女の「ない」のように、マウスを乗せると詳細が現れる。

	A	B	C	D	E	F	G	H
89	「ない」を含む句内の名詞の出現度数と比率〈Y=n〉							
90	No	単語	Group1	比率(%)	Group2	比率(%)	計	比率(%)
91	1	味	0	0	6	11.54	6	10.34
92	2	私	0	0	2	3.85	2	3.45
93	3	商品	0	0	2	3.85	2	3.45
94	4	納得	0	0	2	3.85	2	3.45
95	5	デザイン	0	0	1	1.92	1	1.72
96	6	微妙	0	0	1	1.92	1	1.72
97	7	もの	0	0	1	1.92	1	1.72
98	8	残念	0	0	1	1.92	1	1.72
99	9	評価	0	0	1	1.92	1	1.72
100	10	見た目	0	0	1	1.92	1	1.72
101								
102	「ない」を含む句内の形容詞の出現度数と比率〈Y=n〉							
103	No	単語	Group1	比率(%)	Group2	比率(%)	計	比率(%)
104	1	良い	0	0	2	3.85	2	3.45
105	2	いい	0	0	2	3.85	2	3.45
106	3	よく	0	0	1	1.92	1	1.72
107	4	美味しい	0	0	1	1.92	1	1.72
108	5	まずかっ	0	0	1	1.92	1	1.72
109	6	まずい	1	16.67	0	0	1	1.72
110	7	おいしい	1	16.67	1	1.92	2	3.45
111								
112	「ない」を含む句内の動詞の出現度数と比率〈Y=n〉							
113	No	単語	Group1	比率(%)	Group2	比率(%)	計	比率(%)
114	1	思え	0	0	3	5.77	3	5.17
115	2	思わ	2	33.33	2	3.85	4	6.9
116	3	思っ	0	0	2	3.85	2	3.45
117	4	思う	0	0	2	3.85	2	3.45
118	5	でき	0	0	1	1.92	1	1.72
119	6	言え	0	0	1	1.92	1	1.72
120	7	食べよ	0	0	1	1.92	1	1.72
121	8	いか	0	0	1	1.92	1	1.72
122	9	得	0	0	1	1.92	1	1.72
123	10	られ	0	0	1	1.92	1	1.72
124	11	買い求め	0	0	1	1.92	1	1.72
125								
126	「ない」を含む句内の助動詞の出現度数と比率〈Y=n〉							
127	No	単語	Group1	比率(%)	Group2	比率(%)	計	比率(%)
128	1	ない	2	33.33	10	19.23	12	20.69
129								
130		合計〈名	6	100	52	99.98	58	99.94
131		※四捨五入により100%にならない場合があります。						

図 12 「ない」を含む句の構文解析結果の Excel ファイル
[a] 紙面の都合上、上段部の表示画面を省略した。

表 4 設問内容と回答形式

No	設問内容	回答形式
Q1	名前と学籍番号	記述式
Q2	出身地	選択式
Q3	T 社長の講演をヒントにして、日本農産物の国際競争力を高める方法に関するあなたの考えを自由に書いてください（100 字程度）。	自由記述式

[a]講演とアンケートの実施日は 2014.7.11 の 2 講時である。

ことから、Q2 の出身地属性から図 4 で国外出身（1 人）を除き、東北出身（58 人、6 県）、関東出身（46 人、7 都県）、その他出身（42 人、17 道府県）にグルーピングした。関東出身の都県別平均は 6.6 人、北海道を除いたその他出身の府県別平均は 2.4 人であることから、サンプル数が少ない北海道出身の 4 人は 1 つのグループにせずにその他出身に分類した。Q3 の回答内容に対する出身地の違いを次のように調べた。

まず、東北と関東、東北とその他、関東とその他の出身における図 5 の形態素解析から得られた名詞の中で図 7 のカイ 2 乗検定・フィッシャーの正確確率検定（東北出身と関東出身、東北出身とその他出身、関東出身とその他出身の差）機能を活用し、全体で出現度数 10 回以上の名詞のうち、グループ間に有意差が認められたものを 17 個抽出した。グループ間の違いを調べるために、17 個の名詞を用いた Y=n と Y=1、1≦Y<2 の出現度数をそれぞれ適用し、別途統計解析ソフトの SPSS を用いてコレスポンデンス分析を行った（図 13）。コレスポンデンス分析に用いた数値データは、図 5 で 17 個の当該名詞の□にチェックを入れ、図 5 におけるボタンをクリックすることで取得できた。表 5 に 17 個の名詞について、全体と出身地ごとの出現度数を示した。

図 13 をみると、属性に関わらず言及されやすい原点付近の「競争」を除き、まず、「その他出身」の特徴として、講演者の影響と日本「農業」に対する「中国」の影響を比較的多く取り上げているが、競争力向上に関する具体的な方法を示す単語は含まれていない。「関東出身」は、ほかより「食材」を重要視しており、「文化」の言及率は「その他出身」より高く、図 13 の 1≦Y<2 から「東北出身」より若干高いと考えられた。図 6 から「文化」はユネスコに登録された日本の食文化などを指していることが確認できた。「東北出身」は「付加」「価値」を付けた「販売」に言及しており、有意差のある単語数がほかより 2 倍以上多い。

以上から、日本農産物の国際競争力を高める方法について、「東北出身」はほかより多様な単語を用いながら付加、価値、販売などを、「関東出身」は食材と文化を、「その他出身」は農業と中国をほかより意識している点が出身地の違いとして考えられた。

4.2.3 1≦Y<2 評価法の有効性

図 13 において矢印で示した「文化」は、Y=n（東北出身の出現度数は 5、関東出身の出現度数は 4）では「関東出身」に位置しており、Y=1（東北出身 5、関東出身 3）では「東北出身」に位置している。すなわち、出現度数がそのまま利用される場合と 1 にカウントする場合、その単語のもつ重要度が異なることも生じ、解析に影響を及ぼす。一方、1≦Y<2（東北出身 5、関東出身 3.5）では、「文化」が「関東出身」に位置している。そこで、図 14 に東北出身における 2 回以上出現した名詞（n=168）を対象に 3 つの評価法による出現度数を用いた回帰式を示した。図 14 は図 5 における当該ボタンをクリックし、数値データを

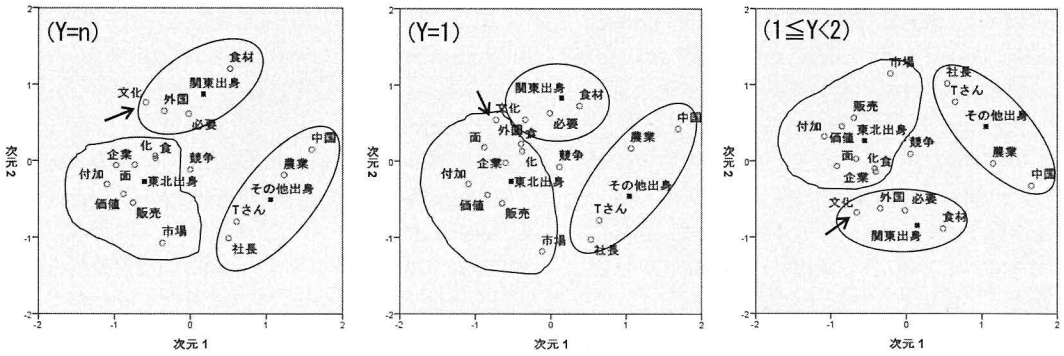

図 13　Q6 の回答文における 3 つ出身地の差

[a]SPSS によるコレスポンデンス分析結果である。

[b]イナーシャの累積寄与率は左からそれぞれ 0.668、0.660、0.676 である。

表 5　グループ間に有意差がある 10 回以上出現単語

No	単語	全体	東北出身	関東出身	その他出身
1	競争	83	41	21	20
2	必要	40	16	17	6
3	農業	40	6	12	22
4	外国	29	14	12	2
5	食	27	16	7	3
6	食材	23	4	14	5
7	化	18	11	5	2
8	社長	17	8	1	8
9	Tさん	17	7	2	8
10	企業	17	12	4	1
11	市場	15	11	0	4
12	価値	15	12	2	1
13	付加	13	11	2	0
14	販売	10	7	1	1
15	文化	10	5	4	0
16	中国	10	0	4	6
17	面	10	7	2	0

[a]表中の度数は Y=n によるものである（紙面の都合上、Y=1、1≦Y<2 は省略）。

[b]各出身地域の度数の中に 0 に近い値があるときはフィッシャーの正確確率検定で有意差を確認した。

[c]全体の数に比べ 3 つの出身地の合計値が不足する数は、国外出身（1 人）によるものである。

ダウンロードして作成したものである。図を見ると、1≦Y<2 はほぼ真ん中あたりに位置し、両極の評価法についていわば中庸的な評価ができることを示唆している。

つまり、分析目的が、当該単語の出現度数そのものに

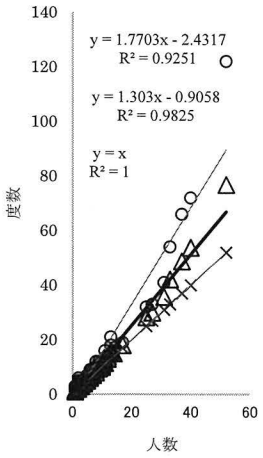

図 14　3 つの出現度数評価法を適用した回帰式

[a]図中の式と線は上から Y=n、1≦Y<2、Y=1 である。

あるときは Y=n を、出現の有無にあるときは Y=1 を用いられるが、本節のように、意図的・習慣的あるいは強調などの理由で繰り返し出現する単語を評価するときは Y=n（過大評価）も Y=1（過小評価）も適切とは言い難く、1≦Y<2 の中庸的な評価法の適用が望ましいと考えられる。なお、分析目的が定かでないときは、図 13 のように、Y=n と Y=1、1≦Y<2 を比較してみることで、Y=n 又は Y=1 の評価法に対し考察を深める契機になろう。

5. おわりに

インターネットユーザが普段慣れているウェブサイト上のボタンをクリックするだけで日本語形態素解析に加え、構文解析やカイ 2 乗検定・フィッシャーの正確確率検定などを順次行うことができるテキストマイニングツールの

iTM を開発した。開発したシステムを用いて事例分析を行った結果、iTM の特徴であるウェブブラウザ上のボタンクリック機能の活用や統計分析用数値データの取得が簡単にできることによって、形態素解析エンジンなどのインストールや煩雑なデータの数値化作業がなくなり、研究者のみならず担い手農業者なども気軽にテキストマイニングを実施できることが明らかになった。また、独自で考案した 1≦Y<2 評価法は、意図的・習慣的あるいは強調などの理由で繰り返し出現する単語の評価法として有効であることが確認できた。

今後、iTM の機能としてコレスポンデンス分析などの統計分析機能を拡張していく予定である。

謝辞

この成果は科学研究費補助金(基盤研究(C):課題番号 25450329)の助成に基づいたものである。

注釈

[注 1] 国立情報学研究所の文献検索サイト CiNii で「テキストマイニング」「農業」などの検索語を用いた文献検索では 21 件が表示された(2015 年 3 月基準)。

[注 2] 意味を持つ最少の文字列の単位を形態素といい、文を単語ごとに分割し、品詞情報などを付け加える作業を形態素解析という(金 2009)。

[注 3] 構文解析とは、文法規則に基づいて、文の構造を句・文節を単位として解析することである(金 2009)。

[注 4] API とは、あるコンピュータプログラム(ソフトウェア)の機能や管理するデータなどを、外部の他のプログラムから呼び出して利用するための手順やデータ形式などを定めた規約のこと。(e-Words 2015a)。

[注 5] ターンアラウンドタイム(Turnaround Time)とは、システムに処理要求を送ってから、結果の出力が終了するまでの時間。データやコマンドの入力が終了してから、処理結果の出力が終わって次の要求の受け入れが可能になるまでの時間のこと。(e-Words 2015b)。

引用文献

石田基広, 2008, R によるテキストマイニング入門, 森北出版, 東京.

e-Words, 2015a, API【Application Programming Interface】, In *http://e-words.jp/w/API.html*, e-Words, 東京.

e-Words, 2015b, ターンアラウンドタイム, In *http://e-words.jp/w/ターンアラウンドタイム.html*, e-Words, 東京.

Google Charts, 2015, Google Charts - Google Developers, In *https://google-developers.appspot.com/chart/*, Google, Mountain View.

堀部 篤, 2014, テキストマイニングによる農業参入法人等と地域との役割分担に関する考察 : 全国農業会議所『解除条件付き農業参入法人等の農地利用に関する調査結果』を手がかりとして, 農村研究, No. 118, pp. 16-28.

飯坂正弘, 2013, 自由記述回答分析による農村移住者の「農村」像 : 描画的なテキストマイニングによる考察, 農村生活研究, Vol. 56, No. 2, pp. 34-41.

磯島昭代, 2006, テキストマイニングを用いた米に関する消費者アンケートの解析, 農業情報研究, Vol. 15, No. 1, pp. 49-60.

磯島昭代, 2009, 農産物購買行動の背景にある消費者の価値意識・ニーズを明らかにするための調査・分析マニュアル—定型自由文へのテキストマイニングの適用—, 中央農業総合研究センター, つくば, p. 9.

IT トレンド, 2015, テキストマイニングソフトを比較して、無料でカンタン資料請求 | IT トレンド, In *http://it-trend.jp/textmining*, IT トレンド, 東京.

金明哲, 2009, 形態素解析と構文解析: テキストデータの統計科学入門, 岩波書店, 東京, pp. 25-36.

KH Coder, 2015, KH Coder Index Page, In *http://khc.sourceforge.net*, KH Coder, 東京.

木南莉莉・木南 章・古澤慎一, 2013, 「戦略的地域農業開発」アプローチの課題 : 新潟県聖籠町を事例に, 新潟大学農学部研究報告, Vol. 66, No. 1, pp. 33-47.

松村真宏・三浦麻子, 2009, 人文・社会科学のためのテキストマイニング, 誠信書房, 東京, pp. 36-37.

三浦麻子, 2012, 社会心理学研究におけるテキストマイニング. 石田基広・金明哲編: コーパスとテキストマイニング, 共立出版, 東京, pp. 141-154.

MeCab, 2015, MeCab (和布蕪)とは, In *http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html*, MeCab, 東京.

Yahoo Japan, 2015, 日本語係り受け解析, In *http://developer.yahoo.co.jp/webapi/jlp/da/v1/parse.html*, Yahoo Japan, 東京.

Technical paper

Development of iTM: a web-based text mining tool

PARK Soo Young[1)], Tadashi HASEBE[1)] and Hiroyuki YASUE[2)]

1) Tokyo University of Agriculture
2) NARO Tohoku Agricultural Research Center

(Received 07 April 2015; in final form 23 December 2015)

Summary

In recent years, there were attempts of study by using text mining in the agricultural sector in Japan as well. The subjects cover a wide range, but there are still only a few papers. The reasons are that the information system is required for morphological analysis of the text, and digitizing for using extracted morphemes in the statistical analysis is a large-scale work. Therefore, in this paper, we have developed a text mining tool, iTM (internet Text Mining), which can sequentially perform Japanese language morphological analysis, parsing, chi-square test and so on, only by clicking buttons on a webpage that is familiar not only with researchers but also with internet users, such as certified farmers. To confirm the utility of the system, we used a web type questionnaire survey tool that is equipped with iTM, and conducted a case study that is targeting, for free, description answer sentences that have been acquired. As a result, it was revealed that we can feel free to carry out the text mining by using the button click function, which is a feature of iTM, since this function makes it possible to eliminate installations such as a morphological analysis engine and the digitizing works of text data. As an evaluation method for word appearing repeatedly, by applying and comparing with a $1 \leqq Y < 2$ valuation method which was uniquely devised by us, it has been suggested that it is possible to deepen the consideration for the analysis results.

Key Words

Agriculture, Frequency of occurrence evaluation, Morphological analysis, Significance test, Statistical analysis

1) 1-1-1 Sakuraoka, Setagaya-ku, Tokyo 156-8502, Japan
2) 4 Akahira, Shimo-kuriyagawa, Morioka, Iwate 020-0198, Japan
(Correspondence: s3paku@nodai.ac.jp)