概要

近年，体験型の観光が注目されている．体験型観光の一つにフルーツ狩りがある．鳥取県においては，梨狩りやブルーベリー狩りなど農業と観光の連携がその一つである．農業従事者や旅行エージェントらは注目の高まっている体験型観光に対して，旅行者により良いサービスを提供するために旅行者の意見や感想を知る必要がある．分析者は旅行者のフルーツ狩りの最中，あるいはその前後の行動を分析することが必要となる．ここで，旅に関するブログ記事に注目すると，旅行者が多くの体験談を記述している．ゆえに，ブログ記事は行動を分析する対象となる．ブログ記事を閲覧すると，ブログ記事にはフルーツ狩りと無関係な記述を見かける．また，ブログ記事の話題に統一性がない．そこで，先行研究 [1] は,「ブルーベリー狩り」に焦点を当て，ブログ記事を類似する体験談にまとめた．さらに，分析に必要な体験談に絞り閲覧するために，動詞を素性に利用したクラスタリングを用いた．しかし，この方法では，ブルーベリー狩りと無関係な記述が混ざることが多く，分析者は余分な文章を読まなければならない．一方，近年，トピックに基づく分類 (LDA;Latent Dirichlet Allocation) が用いられている [2]．LDA は文書素性の次元圧縮に利用可能である．そこで，本研究では，ブルーベリー狩りの行動分析タスクにおいて「LDA におけるトピックを素性に利用する方法」と「動詞を素性に利用する方法」との比較調査を目的とする．

以上について,「動詞を素性に利用する方法」と「LDA におけるトピックを素性に利用する方法」との比較調査を行った．その結果，本研究では明らかにブルーベリー狩りとは無関係なクラスタを 3 件（255 文）にまとめることができた．一方，先行研究では，1 件（78 文）であった．本研究で，トピックを素性に利用することで，分析者の分析効率が高まったことが確認できた．この結果は,「動詞を素性に利用する方法」と比較して,「ブルーベリー狩り」と明らかに無関係な文がまとまっており，分析する際の効率を高めた．また，得られた分析内容を比較した．概ね同様の分析内容を残していることを確認できた．以上により，分析者が読む量を減らせた．また，分析内容は保たれた．

目 次

表 目 次

図 目 次

第1章　はじめに

1.1　背景

近年，体験型の観光が注目されている．体験型観光の一つにフルーツ狩りがある．フルーツ狩りには梨狩りやいちご狩り，ブルーベリー狩りなどが挙げられる．鳥取県においては，梨狩りやブルーベリー狩りなど農業と観光の連携がある．農業従事者や旅行エージェントらは注目の高まっている体験型観光に対して，旅行者により良いサービスを提供するために旅行者の意見や感想を知る必要がある．そのために，農業事業者や旅行エージェントらは旅行者のフルーツ狩りの最中，あるいはその前後の行動を分析することが必要となる．

前述のような情報を得るためには，実際に体験した人が記述した旅に関するブログ記事を閲覧することが挙げられる．旅行者のブログ記事には多くの体験談が記述されており，体験談に対する感想や評価など農業従事者や旅行エージェントらにとって有益な情報が含まれているため，旅行者の行動を分析する対象となっている．しかし，ブログサイトで検索キーワードを打ち込み，ブログを閲覧すると，ブログ記事文中にはフルーツ狩りと無関係な記述を見かける．例えば，ブログ作成者から読者に宛てたあいさつ文などである．この記述はフルーツ狩りとは無関係である．また，ブログ記事の話題に統一性がない．例えば，フルーツ狩りの体験談を記述していたが，途中から連れていたペットの話になるなどである．そのため，分析者はブログ記事を類似する体験談にまとめ，さらに，分析に必要な体験談に絞り閲覧をしたい．

先行研究 [1] では,「フルーツ狩り」のうち,「ブルーベリー狩り」に焦点を当てた．そこで，ブログ記事を類似する体験談にまとめた．さらに，分析に必要な体験談に絞り閲覧するために，動詞を素性に利用することで分析を行った．しかし，この方法では，ブルーベリー狩りと無関係な記述が混ざることが多く，分析者は余分な文章を読まなければならない．一方，近年，トピックに基づく分類 (LDA;Latent Dirichlet Allocation) が用いられている [2]．LDA は文書素性の次元圧縮に利用可能である．そこで，本研究は，ブルーベリー狩りの行動分析タスクにおいて「LDA におけるトピックを素性に利用する

方法」と「動詞を素性に利用する方法」との比較調査を目的とする．

1.2 関連研究

1.2.1 動詞を素性に利用する方法

徳久ら [1] は,「動詞を素性に利用する方法」で類似する体験談をまとめた．ブログ記事中から，体験を表す動詞を検出し，該当する文およびその前後の文を 1 つの体験文章として抽出した．すなわち，3 文を 1 単位とした．また,「ブルーベリー狩り」のフレーズが出現した後半の部分に体験談が多いとし，後半から体験文章を抽出した．前半はブログの読者へのあいさつや経緯説明が多いとし，抽出しなかった．

抽出された「ブルーベリー狩り」に関する体験文章をクラスタリングを行なう際，k-means 法 [7] を用いた．類似する体験談を得るために，ベクトル化に動詞を素性に用いた．抽出した各体験文章中の動詞の有無をベクトルに用いた．また，キーワード抽出のため KeyGraph[3] で処理を行なった．キーグラフに名詞と動詞を文単位で入力に用いることで，クラスタ内の文章全体から主要な単語（動詞と名詞の組）を得る．

分析者が，文章を閲覧する際にはポジティブな情緒の推定された文章を中心にするとした．

結果として，元となる記事は,「ブルーベリー狩り」または「ブルーベリー摘み」の表現を含む文，642 件，15,328 文である．これから，記事の前半から文を閲覧対象から削除した．10,897 文となった（4,431 文削減，圧縮率 71%）．体験動詞に基づき，体験文章を 1,382 文章，4,146 文を得た．全てのクラスタからポジティブな感情の推定される文章を閲覧することにすると，1,068 文章，3,204 文を得た（圧縮率 21%）．また，分析者は 414 文書 1,242 文を熟読した．

1.2.2 要素技術について

Latent Dirichlet Allocation

LDA とは，文書生成モデルである．すなわち，複数のトピック $\mathbf{z}$ からある確率で文書 $\mathbf{w}$ が生成されるとするモデルである．$\mathbf{z}$ と $\mathbf{w}$ の生成確率は次式で求められる．

$$p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta), \tag{1.1}$$

$\mathbf{z}$ はトピックベクトルを表す．また，θ はトピックの混合比，α および β はパラメータである．なお，w_n は n 番目の単語である．

トピックとは話題を支える単語の集合で，トピックごとに何らかの共通の話題に対する単語が集まっている．

このモデルを用いると任意の文書をトピックによるベクトルで表すことができる．そのためのトピックは，大量の文書から学習しておくこととする．本研究では，[11] のツールを利用する．

以下に本研究を進める際に参考にした LDA を用いた研究を述べる．落合らは LDA の確率を割り当てる対象を述語項構造を基本とした単語の組にすることで，商品に対するレビュー文書の動詞による特徴を抽出した [4]．芹澤らは LDA を用いて，トピックを抽出し，文書内の語の特徴量を term-score で計算した．各トピック間の類似度をコサイン類似度で測った [5]．立川らは LDA でトピックを抽出する際に，与えられている文書から制約となる単語群を自動抽出し，事前知識として与えることで制約を踏まえたトピック抽出を行った [6].

Yahoo!LDA

本研究では，LDA を用いたベクトル化において Yahoo!LDA[11] というツールを用いる．Yahoo!LDA の主な出力としては表 1.1 のファイルがある．

表 1.1: Yahoo!LDA の出力

出力ファイル	概要
lda.docToTop.txt	トピック番号とそれに帰属する量
lda.worToTop.txt	ID ごとの単語とそれが属するトピック番号
lda.topToWor.txt	トピックごとに属する単語とそれの量

lda.docToTop.txt はベクトル化に用いる．lda.worToTop.txt は式 (1.1) で各単語に割り当てられたトピックを θ によって調整した値を出力している．

クラスタリング

クラスタリングとはある事例集合について，類似する複数の事例をまとめていくつかの部分集合にすることをクラスタリングという．その部分集合のことをクラスタという．

クラスタリングには階層型クラスタリングと非階層型クラスタリングがある．階層型クラスタリングには凝集法がある．N 個の事例が与えられたとき，1 個の対象を含むクラスタから始めて，クラスタ間の距離から逐次的に併合する方法である．一方，非階層型クラスタリングには k-means 法がある．本研究では，k-means 法を用いることとする．

k-means 法はランダムでクラスタに事例を割り振り，割り振った事例をもとに各クラスタの中心を計算する．計算は割り当てられた事例のベクトルの平均を用いる．全ての事例において，事例の属するクラスタの平均とそのデータとの距離が最小になるように，事例の属するクラスタを決め直すものである．

本研究では，クラスタリングのツールに，k-means 法に対応している bayon[7] を用いる．大規模なデータに対して，高速に実行可能である．

キーワード抽出

キーワード抽出とは，対象文書中での登場頻度，互いの繋がりの強さで，重要な単語を抽出することである．一つのキーワード抽出における手法として，TF-IDF がある．TF-IDF は単語の特徴度を計算する．しかし，行動分析のためにキーワードを得るためには，対象と行動の組を得る方が，分かりやすい．ゆえに，共起に注目する．

共起を利用するキーワード抽出に KeyGraph[3] を用いた研究がある．KeyGraph は文書の単語の頻度，および，単語間の共起関係について「土台」,「屋根」,「柱」という考えを用いた．KeyGraph は，これらの考えのもと，文書の主張とその関連語を抽出できる．本研究では，KeyGraph を用いたキーワード抽出を用いる．

情緒推定

本研究では，文章の情緒推定に，パターン辞書を用いる [8],[9].

この辞書は，日本語語彙大系 [10] の「結合価パターン」に情緒属性を付与した辞書である．情緒原因表現，情緒の状態表現，および情緒の表出表現が判定できる.「情緒主」,「情緒対象」,「情緒名」が付与されている．情緒名は，≪喜び≫，≪悲しみ≫，≪好ましい≫，≪嫌だ≫，≪驚き≫，≪期待≫，≪恐れ≫，≪怒り≫，≪なし≫の 9 種類が用いられる．パターン数は，約 14,800 パターン収録されている．

この辞書を利用した情緒推定ツール patlap（本研究室で作成されたツール）を用いて，情緒推定を行う．本研究においては，patlap によって Positive な情緒の推定された文を分析対象とする．

1.3 本研究の目的

本研究では,「ブルーベリー狩り」の観光開発を想定し，[1] との比較調査を行なう．

比較調査において，ベクトル化の方法が，[1] と本研究とは異なる．

[1] では，体験文章における動詞のみをベクトル化に用いた．一方，本研究では，体験文章におけるトピックを LDA を用いて取得し，それらをベクトル化に用いる．LDA を用いることで，動詞以外の単語についてもトピック分類に重要な単語であれば参照されるため，分類性能の向上が期待できる．

以上のベクトル化の相違により,「動詞を素性に利用する方法」と「LDA におけるトピックを素性に利用する方法」の比較調査を目的とする．

本研究における評価としては,「ブルーベリー狩り」とは無関係な記述をまとめることで分析者が読み捨てられる分量，および,「ブルーベリー狩り」の分析を実際に行なって先行研究と本研究の分析内容の一致で評価する．

第2章　ブログ記事の分類方法と行動分析

本章では，ブログ記事の分類方法と行動分析について説明する．

2.1　分類と分析の主旨

本研究における分類と分析について述べる．

分類とは，ブログ記事中の多数の体験談を対象に,「ブルーベリー狩り」の体験中，または体験前後の行動を中心とした類似する体験文章に分類することである．また，ブルーベリー狩りと無関係な記述をまとめた分類を得ることである．

分析とは，ブログ記事中にある多数の体験談から「ブルーベリー狩り」の行動タスクを分析することである．分析者はブログ記事中のブルーベリー狩りの体験中，または体験前後の行動を中心とし，分析対象とする．

2.2　基本的な流れ

ブログ記事の分類方法は，体験文章の抽出，体験文章のベクトル化，体験文章の分類，体験文章の情緒推定，体験文章からのキーワード抽出および体験文章の分析を一連の流れとする．本節では，まず，図 2.1 にブログ記事の分類と分析の一連の流れを図で表す．次に，分類手順，および行動分析手順を説明する．

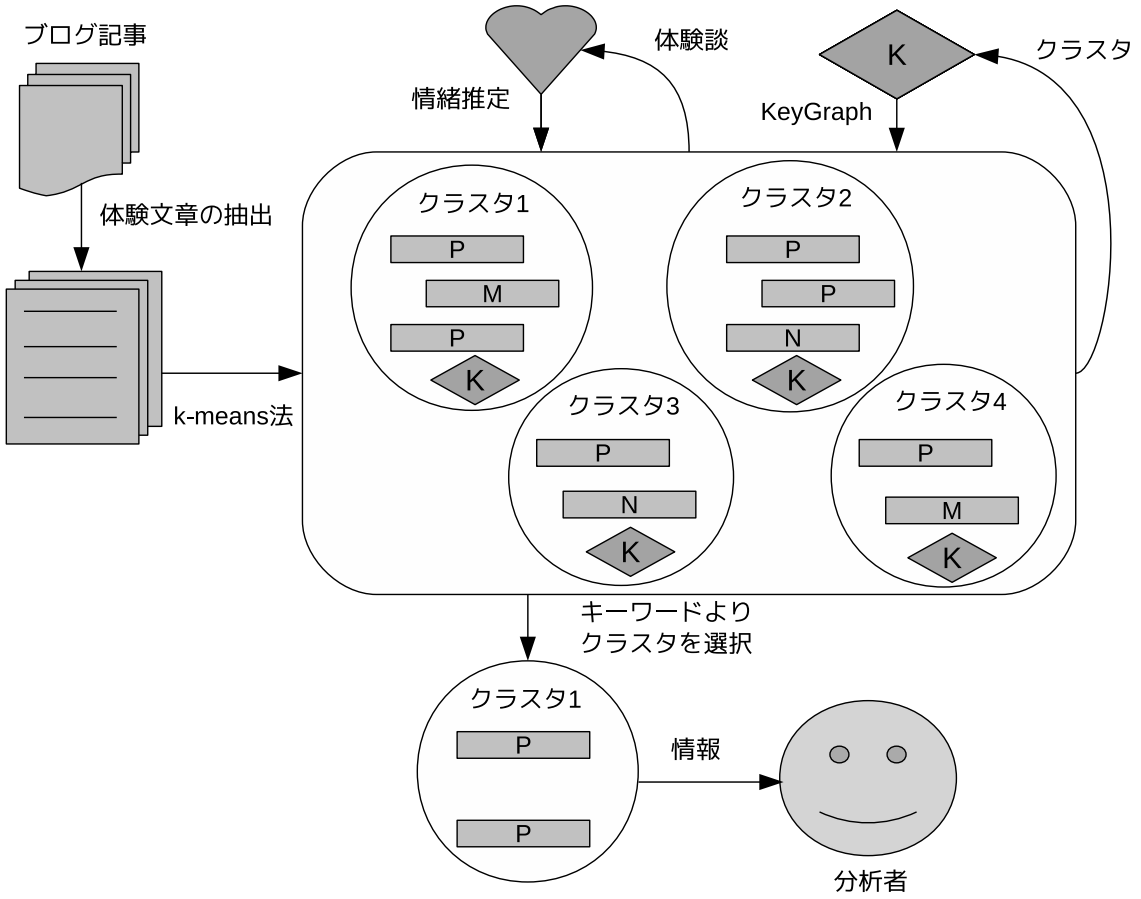

図 2.1: ブログ記事の分類と分析の一連の流れ

ブログ記事を以下の 1〜3 で分類し，4〜7 で分析する．

1. 「ブルーベリー狩り」のブログ記事において「ブルーベリー狩り」または「ブルーベリー摘み」の表現を含む文の 1 文前から記事の末尾までの範囲より，体験文章 (体験動詞を含む文+その前後 1 文) を抽出し，体験文章集を作成する．
2. 体験文章集に LDA におけるトピックを用いることで，ベクトルを付与し，ベクトル文章を作成する．
3. 2 で作成したベクトル文章を用いて，k-means 法によりクラスタリングを行なう．
4. クラスタごとに，KeyGraph でキーワードペアを生成する．
5. ブログ記事の分析者はクラスタのキーワードペアをみながら，閲覧するクラスタを選択する．
6. ポジティブな情緒の推定される文章を表示する．
7. 分析者は 6 で表示された文章を熟読し，行動を分析する．

2.3 分析の手順

具体的な分析の手順を説明する．

分析者は，クラスタごとにキーワードを読む．クラスタに対して,「期待あり」,「期待なし」,「曖昧」の 3 択で評価する.「期待あり」のクラスタは，ポジティブな文を熟読する.「期待なし」および「曖昧」のクラスタは，斜め読みとする．

斜め読みの過程で，興味がわけば熟読することとする．逆に,「期待あり」のクラスタでつまらないと判断した場合，斜め読みに切り替えることを認める．熟読したか，斜め読みしたかは記録する．

熟読することで，体験談を得られる．発見した体験談はメモをとることとする．類似の体験談は読み飛ばすことを認める．また，斜め読みの途中でも気になる文章があれば，その文章だけは熟読し，体験談のメモをとる．

以上より，各クラスタに対して，期待したかどうか，熟読したか，期待正しさ，および体験談のメモを残す．

第3章　実装

本章では，本手法の個々の処理部および実装について説明する．

3.1　システムの概要

3.1.1　システムの構成

本システムの構成を図 3.1 に示す．本システムは「体験文章抽出部」,「ベクトル化部」,「クラスタリング部」,「キーワード抽出部」，および「情緒推定部」から構成する．

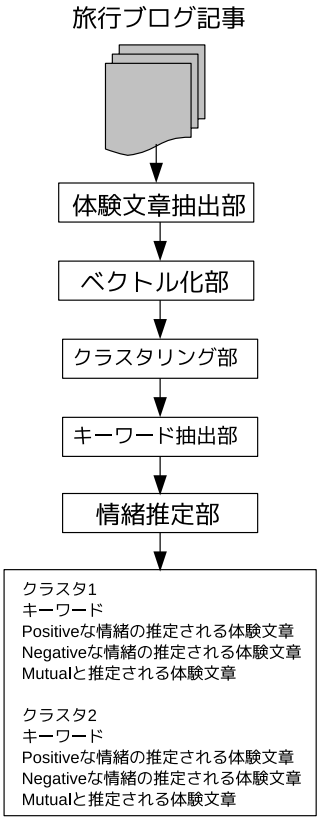

図 3.1: システムの構成

3.1.2 システムの流れ

本システムの流れを以下に示す.

1. 体験動詞をあらかじめ辞書に登録し，該当文および前後 1 文を体験文章として出力する.
2. LDA におけるトピックを素性とし，1 で抽出された体験文章をベクトル化する.
3. 2 でベクトル化された値をもとに，k-means 法でクラスタリングを行なう.
4. KeyGraph を用いて，3 で生成されたクラスタごとに動詞および名詞の組のキーワードを抽出する.
5. 3 で生成された各クラスタに結合価パターン辞書を用いて，Positive，Negative, および Mutual の 3 分類の情緒推定を行ない，出力する.

3.1.3 システム環境

本システムは次の環境上に実装する．OS には VineLinux5.2 を，プログラミング言語には Ruby を用い，ツールには MeCab（形態素解析 [12]），Yahoo!LDA（トピックモデルを行なうツール [11]），KeyGraph（キーワード抽出ツール [3]），patlap（本研究室で作成された情緒推定ツール [8]，[9]）をそれぞれ用いる．本研究は以上のものが使用できる環境で行なう.

3.1.4 体験文章抽出部

準備

ブログ記事から体験文章を得るために，体験動詞を基準とした．体験動詞は，一般ブログ記事から「〜してみる」を基本とする動詞抽出し，あらかじめ辞書に登録しておく．表 3.1 に辞書に登録されている動詞の抽出ルールを示す.

表 3.1: 体験表現抽出ルール

条件	「V+てみる」「V+てみた」「V+てみます」 「V+てみました」「V+たことがある」 「V+たことがあった」のいずれかにマッチ
出力	条件より抽出された，V の標準系

実行

2.2 節で説明した流れ 1 において，体験動詞を含む文（以下体験文と呼ぶ）の前後 1 文ずつを追加した 3 文を体験文章として抽出する．また,「ブルーベリー狩り」に関する体験文章を得るために，ブログ記事において,「ブルーベリー狩り」と記載された文より後半の文から体験文章を集める．すなわち，記載された文より 1 つ前の文から記事末までである．理由として，前半にはブログ作者の読者への挨拶などが含まれるためである．

3.1.5 ベクトル化部

準備

ベクトル化の準備として，事前に Yahoo!LDA というツールを用いて，トピックモデルを一般ブログ記事より作成しておくこととする．図 3.2 に，一般ブログ記事より作成されたトピックモデルのうち，例としてトピック番号 1 を示す．また，図 3.3 および図 3.4 に「ブルーベリー狩り」のブログ記事より，作成されたトピックモデルのうち，例としてトピック番号 4 および 41 を示す.

Topic 1: (見,0.14593) (テレビ,0.103908) (い,0.0950574) (し,0.092951)
(番組,0.0796922) (放送,0.0696462) (ニュース,0.0442126) (ＴＶ,0.0406479)
(ドラマ,0.0380889) (いる,0.0310098) (ＮＨＫ,0.0302443) (さん,0.0294677)
(ラジオ,0.0278417) (て,0.0278362) (観,0.0265678) (の,0.0261935)
(やっ,0.0240312) (今日,0.0239418) (出演,0.0217516) (最近,0.0209805)

図 3.2: 一般ブログ記事よりトピック例

Topic 4: (今日,0.159098) (お,0.114063) (行き,0.0771337) (来,0.0709762)
(昨日,0.0682837) (行く,0.0585112) (さん,0.0564674) (帰り,0.0417956)
(仕事,0.0360259) (友達,0.0355891) (明日,0.0353544) (家,0.0331346)
(い,0.0296011) (また,0.0292654) (久しぶり,0.0281603) (後,0.0275475)
(一緒,0.0269347) (朝,0.0249463) (前,0.0236294) (買い物,0.0234827)

図 3.3: ブルーベリー狩りブログ記事よりトピック番号 4

Topic 41: (パン,0.111392) (味,0.08301) (ケーキ,0.0753528) (美味しい,0.062648)
(お,0.0555046) (コーヒー,0.0537515) (い,0.0493889) (アイス,0.0492076)
(の,0.0481396) (おいしい,0.0466988) (もの,0.0465578) (食べる,0.0453085)
(焼き,0.0375002) (好き,0.0374699) (み,0.0371878) (チーズ,0.0360493)
(ジュース,0.0322006) (作り,0.0316364) (野菜,0.0309915) (デザート,0.0300042)

図 3.4: ブルーベリー狩りブログ記事よりトピック番号 41

実行

抽出された体験文章をベクトル化する．本システムでは，ベクトル化に 1.2.2 節で説明した LDA におけるトピックを用いる．体験文章をベクトル化するために，Yahoo!LDA を用いる．本システムでは入力として，Yahoo!LDA の出力ファイルである lda.docToTop.txt を用いる．以下にベクトル化の様子を示す．図 3.5 に 3.1.4 節で抽出された体験文章の例文を挙げ，図 3.6 にベクトル化の例を挙げる．

今日 ゴルフ 以外 話題 先週 皆 連れる * 遠征 ラウンド 行く くる の その 帰り道 大きい ＮＺ 有名 ブルーベリー 農園 ある ブルーベリー 狩り 行く くる ブルーベリー 健康 良い 特に 目 良い こと 受講 生 達 摘み たて ブルーベリー 喜ぶ 食べる いる

図 3.5: 体験文章例

図 3.6 に示す例では，括弧の 1 番目の要素はトピック番号を指し，2 番目の要素はそれに対応する量を示している．

3.1.6 クラスタリング部

クラスタリング部では，抽出された体験文章集を 24 のクラスタに分割するとともに，「ブルーベリー狩り」のブログ記事の類似する体験談をまとめる．

(41,0.275) (4,0.125) (27,0.1) (57,0.1) (38,0.075)
(60,0.05) (65,0.05) (81,0.05) (84,0.05) (29,0.025)

図 3.6: ベクトル化例

多くの関連行動を読むことは，分析者にとって困難である．そこで，類似する体験文章を先行研究と同様の 24 のクラスタに機械的に分割する．本研究では，bayon[7] を用いる．k-means 法を用いる．前節で説明した LDA におけるトピックを素性としたベクトルを入力とし，出力は 24 のクラスタとする．

3.1.7 キーワード抽出部

前節で得られたクラスタの概要を把握するため，また，行動分析の手がかりとするためにキーワード抽出を行なう．本システムでは KeyGraph[3] を用いる．KeyGraph はその文章の主張と関連語を抽出することができる．

入力として，得られた 24 のクラスタを用いる．各クラスタごとに文単位で処理をし，動詞および名詞の組を出力する．表 3.2 にキーワード抽出を行ない，各クラスタごとに得られた動詞と名詞の組を一部示す．

表 3.2: 動詞および名詞のキーワードペア

♯	動詞	名詞
1	Y:する (10)	T:鍋 (1) T:重さ (1) T:水 (1) T:下ろし (1) T:レモン (1) T:掏摸 (1) T:薫り (1) T:豆腐 (1) T:硬め (1) T:出来上がり (1)
2	Y:食べる (14)	T:木 (1) T:写真 (1) T:ん (1) T:たくさん (1) T:お (1) T:ー (1) T:ベリー (1) T:ブルー (1) T:パック (1) T:立て (1) T: ”(1) T:食べ (1) T:後ろ (1) T:ジャム (1)
3	Y:する (9)	T:森 (1) T:景色 (1) T:ヶ (1) T:駅 (1) T:ベリー (1) T:ブルー (1) T:w (1) T:道 (1) T:牧場 (1)

表 3.2 の動詞の前の Y は用言を表し，T は体言を表す．また，数字は出現回数を示す．

3.1.8 情緒推定部

本システムにおいて，情緒推定を行うために,「結合価パターン辞書」を用いる．1.2.2 節で説明したように，日本語語彙大系の結合価パターンに情緒属性を付与して作成された辞書を表す.

各クラスタの体験文章を入力とする．出力は Positive，Negative，Mutual の 3 つに分類した体験文章とする．本研究では，ポジティブな情緒は +1，ネガティブな情緒は −1 とし，さらに複数が推定される場合は平均化することで，Positive，Negative，および Mutual の 3 分類の情緒をあつかう．表 3.3 にクラスタ 1 の Positive，Negative，Mutual の 3 分類を行なった例を示す.

表 3.3: 情緒推定後の出力例

推定される情緒	体験文章
Positive	こんなに取れた。お砂糖を入れて煮詰めます。 美味しいジャムになりました。
Negative	（当時作ったケーキは、ゼラチンの粒が 残ってしまって口当たりが悪かった） 今回のケーキは大人になった私には難易度は低く、 ふんわり美味しく出来上がった。ブルーベリージャムは少し水でのばして、砂糖も少し加えレンジで加熱してソースに。
Mutual	★作り方★タッパーなどに水気をきったブルーベリーを入れ、 やや多めの砂糖を入れて軽くふる。そのまま冷凍庫へ。

第4章　実験

「ブルーベリー狩り」の観光振興考案を獲得するために，分析者がブルーベリー狩りの観光体験談を分析するという状況を想定する．

4.1　実験条件

トピックの学習は，ある1つのサイトの2009年6月，7月および8月の金曜日，土曜日，日曜日の一般ブログ記事 (9,642,782文) から行なう．トピック数は100とする．ブルーベリー狩りのブログ記事は，あるブログサイト3つから2010年に書かれた「ブルーベリー狩り」または「ブルーベリー摘み」の表現を含む記事である．15,328文を用いる．なお，k-means法によるクラスタ数は，先行研究と同じく24とする．

4.2　分類結果

文の量を示す．元となるブログ記事は，15,328文である．抽出された体験文は4,146文である．2.2節の流れ3において，24クラスタを得た．

表4.1に各クラスタに含まれていた文章数および文数を示す．表4.1をみると，クラスタ2，クラスタ10に文が集中している．

表 4.1: 各クラスタの文章数および文数

♯	文章数	文数
1	60	180
2	233	699
3	66	198
4	31	93
5	40	120
6	39	117
7	47	141
8	43	129
9	35	105
10	159	477
11	26	78
12	52	156
13	30	90
14	29	87
15	47	141
16	54	162
17	41	123
18	35	105
19	38	114
20	34	102
21	40	120
22	62	186
23	46	138
24	95	285

2.2 節の流れ 3 において，k-means 法によってクラスタリングを行なった結果，および動詞と名詞のキーワードの組の一部を表 4.2 に示す.

まず，クラスタリングを行なった結果，表 4.2 を見ると，クラスタ 1 には料理に関する文が集まっている．次に，クラスタ 2 を見ると，ブルーベリーの味であったり，おいしい食べ方が集まっている．最後に，クラスタ 3 において，ブルーベリー狩りの前後の道中での行動が集まっている.

表 4.2: クラスタリング結果

♯	例文／主なキーワード
1	・山ほどのブルーベリーを大きさと熟れ具合で選別して、生食用とジャム用に分けた。 ジャム用は早速砂糖とレモン汁を入れて煮た。 良く熟れていたので、砂糖の量が少なくても大丈夫だった。 ・去年頂いた時はジャムにしたので、今年はサワードリンクにしてみました！ 漬けたそばからブルーベリーの色がお酢にうつってきてました。 約２週間後に完成です！ ・まずは、ジャム作り軽く洗ってお砂糖を混ぜる 弱火で煮始め少し火を強めて混ぜ混ぜ ／（動詞）煮る （名詞）鍋 重さ 実 レモン 硬め
2	・早速、私の大好きな“ 胚芽食パン ”を焼きスライスしたパンの上に ブルーベリージャムとクリームチーズを乗せていただきました チョーーーー絶品 ・甘くて美味しいんですよ。 帰りにはブルーベリーのアイスクリームも食べて、練習の合間の良い リフレッシュでした。そしてこれが摘んだブルーベリーです。 ・広い敷地に１０年程の立派なブルーベリーの樹々が目の前に広がります。 勧められていただいた白く粉を吹いたような完熟の摘みたてブルーベリー。 甘くてほんのり酸っぱい…ｙｕｍｍｙ（ｏ＾－＾ｏ） /（動詞）持ち帰る （名詞）って たくさん ベリー ブルー 土産 味
3	・は爺婆には堪えるのでパスさせて貰った。 最後は、大山の南山麓の“ ブルーベリー狩り ”に向かった。先ず、途中の “ 鬼女台展望休憩所 ”で、烏ヶ山（手前）と大山（遠方）の景色を楽しんだ。 ・お友達のところでブルーベリー狩りをさせていただきました。 帰り道「馬滝」という素朴な標識が目に入ったので行ってみました。 たくさん曲がって細い山道をずっと上っていくとありました小さな素朴な滝が。 ・でも、「雨天につき閉園」ということで … 残念！ しかし、タダでは終わらないのが、この「瑞香園」です。 近くの県道沿いに、関連のお店があって、そこでケーキなどなどが楽しめます ／（動詞）する （名詞）森 景色 ヶ 駅 ベリー ブルー ｗ 道 牧場

4.3 分析結果

2.2 節の流れ 3 において，すべてのポジティブな情緒の推定される文を読むこととすると，3,204 文となった．分析者のクラスタの閲覧結果を表 4.3 に示す．また，表 4.4 にクラスタ中のポジティブな情緒の推定される文章数および文数，分析者が熟読した文章数および文数を示す．また，分析者がクラスタのキーワードにより，明らかにブルーベリー狩りと無関係なクラスタであると判断した，クラスタの一部を表 4.5 に示す．

表 4.3: クラスタの閲覧結果

♯	期待	熟読	期待の正しさ	メモ数
1	曖昧	した	一致	7
2	あり	した	一致	42
3	曖昧	した	一致	9
4	曖昧	せず	一致	2
5	曖昧	せず	一致	1
6	曖昧	せず	一致	2
7	曖昧	せず	不一致	4
8	あり	した	一致	4
9	曖昧	せず	不一致	1
10	あり	した	一致	11
11	あり	せず	不一致	2
12	なし	せず	一致	0
13	あり	せず	不一致	0
14	なし	せず	一致	0
15	曖昧	せず	一致	3
16	あり	した	一致	7
17	曖昧	した	一致	2
18	曖昧	せず	不一致	1
19	あり	せず	不一致	2
20	なし	せず	一致	0
21	曖昧	した	不一致	3
22	曖昧	せず	一致	3
23	あり	した	一致	2
24	あり	した	一致	7

熟読された文のうちのメモの総数は 345 文である．

クラスタ 12，14 および 20 は，分析者が興味を持てず，かつ斜め読みを行なう過程で内容に興味が持てないと判断した．ゆえに，期待の正しさが一致した．

表 4.4: クラスタの分析結果

♯	文章数	文数	熟読文章数	熟読文数
1	41	123	41	123
2	206	618	206	618
3	53	159	53	159
4	23	69	0	0
5	28	84	0	0
6	31	93	0	0
7	35	105	0	0
8	34	102	34	102
9	26	78	0	0
10	116	348	116	348
11	19	57	0	0
12	38	114	0	0
13	21	63	0	0
14	24	72	0	0
15	40	120	0	0
16	42	126	42	126
17	28	84	28	84
18	29	87	0	0
19	23	69	0	0
20	23	69	0	0
21	28	84	84	252
22	43	129	0	0
23	33	99	99	297
24	83	249	249	747

熟読文数において，クラスタ 2 に期待を持ち，期待通りであったため分類を行われた中で，文数の一番多い，クラスタ 2 が熟読文数が多いという結果となった．

無関係なクラスタとして，クラスタ 12 において，ブルーベリー狩りの話題とは異なり，

表 4.5: クラスタの分類結果

♯	例文／主なキーワード
12	・二人ともジージのベルトを握りしめ、緊張してるのがよくわかる。 それでもショベルがガバッと土を掘り上げるところは興味津々の様子。 「石投げ！」と「ユンボー！」。 ・途中タングラムというスキー場のところで小休憩です。 ここでＡＫＢ４８さんが膝の古傷の痛みを訴えていました。 テーピングを巻こうということになったのですが、、、、、 テープを剥がすとき、毛が引っ張られて痛いということで、急遽ハサミでカット。 ・あとは、いろいろ想い込めて手紙書こう。 花束もホワイトデーのとき喜んでくれたから、 またちゃんと用意しよーうひょーーーーーーたーのしーみだーーーーーー ／（動詞）掘る（名詞）津々 ガバッ ショベル 興味 ところ 土 様子
14	・↑いつも応援ありがとうございます。 クリックしていただけると、嬉しいです。 みなさんの応援が次回の更新の励みです！ ・ ｉｎｆｏ@ｔｕｎａｇｕｋａｉ．ｃｏｍ 注文してくださった方の声はこちらをクリック 詳しいセット内容を確認したい方はコチラをクリック ・にほんブログ村何時も応援ありがとうございます。 ↑↑２つもお願いして申し訳ないですが ‥ こちらもポチっとしていただけると 喜んで舞い上がります。豚もおだてりゃ木に登るの心境です。 /（動詞）お願いする （名詞）ー 上 ポチ リンク っ 予約
20	・自己嫌悪今日は予約がなかったから良かったものの一人で経営するということは なんでもあり～なんだけど、自分をきちんと管理しないといけないんだ !! と、じゅんこさんは強く反省しました。 ・メールをすべきではないというより、 会話が一段落するタイミングに合わせたり、 一言ことわるなど、配慮が大切だと思われます。 ・肉の取扱い業者の方々は、やはり今話題である『口蹄疫』の話に触れていました。 魚の業者は、ノルウェー産の魚を買い取り自社で加工している努力を 話してくれました。→市場に出ているノルウェー産の切身などは、 中国で加工されているものがほとんどであるのが現状 ／（動詞）思う （名詞）気持ち って 感謝 が 恐れ

祖父と孫でシャベルを使って何かを掘る様子を表す文，怪我をしたときの様子が記述されている．また，クラスタ 14 において，外部サイトへアクセスをうながす文が集まっている．クラスタ 20 のおいて，当時話題になっていた「口蹄疫」や男女関係のアドバイスをする記述がまとまっていると分析した．

4.3.1 得られた体験談の集計

分析者は 24 クラスタにおいて，熟読した文のうち 345 文のメモを作成した．体験文章の閲覧時間はメモの作成を含め，約 4 時間 45 分を要した．また，体験談のメモを手作業で分類すると，次に示す A から J の項目を挙げることができた．これには約 2 時間を要した．

メモの内訳・分類を以下に示す．

A. 同行者

- 家族，子連れ
- サークル仲間

B. 移動手段

- 車
- 自転車
- バイク

C. ブルーベリー狩り最中の体験

- 太陽の当たっている所が，甘味が強い
- 大きい実の方が断然おいしい
- 実が赤いのが酸っぱく，黒いのがおいしい確率が高い
- ブルーベリーの育て方を農園のご主人から教えてもらう
- ブルーベリーが暑さで暖かい
- 急に雷雨にあう

- 虫にさされる
- カメラで撮影する
- 必要なものは貸し出しを行なっている

D. ブルーベリー狩り前後の農園での体験

- 園内でジェラートなどを食べる
- ブルーベリー狩りとともに，とうもろこし，ラズベリー，ズッキーニ，トマト，イチジク，柿の収穫
- 実の摘み方の講習を受ける
- ジャムづくりを体験する
- 野菜畑を見学し，野菜ソムリエから収穫の仕方やおいしい食べ方のレクチャーを受ける

E. ブルーベリー狩り前後の近隣施設での体験

- 近くの木陰で一休み
- 園内の休憩所で休憩
- ショッピングをする
- バーベキューをする
- レストランなどで昼食をとる
- カフェで一休み
- 温泉や露天風呂に行く
- 紫陽花の鑑賞
- ブルーベリーのタルト，コーヒー，ソフトクリームを食べる
- 直売所でカスタードクリームとブルーベリーのデニッシュを食べる
- 土産ものを売る店がある

F. ブルーベリー狩り前後の道中の体験

- ラベンダー畑へ行く
- 農作物の直売所に立ち寄る
- 展望台，高台で景色を楽しむ
- 山歩き，登山を行なう
- 蛍鑑賞
- スイカ割り
- ダムに立ち寄る
- 滝を鑑賞
- 乗馬体験をする
- カブトムシ採りをする
- スイーツガーデンに立ち寄る
- プールで遊ぶ
- 川遊び
- ボート遊び

G. ブルーベリー狩りを包含するイベント

- ゴルフをする
- 仕事の流れで行く
- 講演会に参加する
- ハイキングをする
- キャンプをする
- 花火大会を鑑賞する

- 部活での練習の合間に行く
- サークル活動で訪れる
- バスツアーで訪れる
- スクール合宿の一環で訪れる
- ドライブ

H. 帰宅後の体験

- ジャムをつくる
- ヨーグルトにかけて食べる
- パンをつくる
- ジャムとバターで食べる
- 酢を使ったサワードリンクつくり
- 胚芽食パンにブルーベリージャムとクリームチーズをのせる
- にんじんベーグルにジャムをのせる
- ジャムをアイスクリームバニラ味にかけて食べる
- ケーキを焼く

I. ブルーベリー狩りにおける注意事項

- 日焼け止め，虫除けスプレーは必須
- 炎天下の中の作業である
- 帽子があった方が良い
- 人気のある予約制の農園では断られる場合もある

J. その他

- ブルーベリーの大きさや熟れ具合で生食用やジャム用に選別する

- ブルーベリーを持ち帰ることが可能である
- 目に良い
- 他のフルーツ狩りより安価である
- 1 時間も経たないうちに終了
- 小さい子には採っていいものとの区別がつかない
- 練馬区では観光農園を PR している
- 練馬区の西側の地区にブルーベリー園が多い

第5章　考察

考察として，本研究と先行研究との読み捨てる量と分析内容の比較を行なう．

5.1　読み捨てる量の比較

4.3 節において，クラスタを分析した結果を示した．表 4.3 より，クラスタ 12，14，および 20 の 3 件のクラスタを事前にキーワードを読むことで「ブルーベリー狩り」の体験談から，不要なクラスタであると判断した．その文数は表 4.4 より，255 文である．一方，先行研究の手法ではクラスタ 1 の 1 件のクラスタ，78 文だけであった．ゆえに，分析者の体験談を読む上で，読み捨てる量が増加した．つまり，分析する効率が高まったと考えられる．

閲覧にまったく適さない文章としては，表 4.5 より，ブログに訪問した読者に対して作者が外部へのサイトへ誘導する文である．例えば「こちらをクリック」という文であった．また，最近流行している話題をとりあげたブログ文である．例えば「口蹄疫」という文であった．先行研究においては，本研究と同様に「こちらをクリック」のような表現を含む，ファシリエイトなどを誘う文である．比較すると，先行研究にも見られたファシリエイトを誘う文に加え，流行や男女関係などを話題にしているような文をまとめていると分析した．

5.2　分析内容の比較

分析内容の比較を行なった．前節において，読み捨てる量を述べた．読み捨てるクラスタ内に，ファシリエイトなどを誘う文が先行研究と同様に分類されていた．また，読み捨てられるクラスタ以外の調査を行なった．読み捨てられるクラスタ以外から得られた分析内容を比較対象として，内容に差があるかを対象とした．結果として，分析者により表現の仕方が違うが，概ね同様の分析内容を残している．

第6章　おわりに

近年，体験型観光が注目を集めている．そこで，本研究ではフルーツ狩りに着目して，その前後の行動を分析することで，農業従業者や旅行エージェントがより質の良いサービスの提供や観光案内ができると考えた.「ブルーベリー狩り」の観光振興案を獲得することを想定して，その観光振興案の体験談をまとめ，分析するために「LDA におけるトピックを素性に利用する方法」を提案した.

トピックを素性に利用することで，分析者の分析効率が高まったことが確認できた．この結果は,「動詞を素性に利用する方法」と比較して,「ブルーベリー狩り」と明らかに無関係な文がまとまっており，分析する際の効率を高めた．また，分析内容を比較した．概ね同様の分析内容を残していることを確認できた.

今後の課題として，より分析効率を向上させるとともに，不要なブログ記事を学習する．また，SPAM のメールフィルタと同様な手法で削除をするという方法を行うことで分析者が事前に分析対象と無関係な文章を削除することである.

謝辞

本研究を進めるにあたり，御指導，御助言を頂きました，鳥取大学工学部知能情報工学科計算機工学講座 C の村田真樹教授に心から御礼申し上げます．本研究を進めるにあたり，御指導，御助言を頂きました，村上仁一准教授に心から御礼申し上げます．また，徳久雅人講師には，終始にわたり研究の進め方や論文の書き方など細部にわたる御指導を頂きました．ここに深く感謝いたします．その他様々な場面で御助言を頂いた計算機工学 C 講座の皆様に感謝の意を表します．

参考文献

[1] 徳久雅人, 山本拓未, 村田真樹, 村上仁一: “ブログ記事からの観光体験談の抽出 — クラスタリングとキーワード抽出を用いる場合”, 観光情報学会 第 6 回研究発表会, pp.89-94, 2012.

[2] David M. Blei, Andrew Y. Ng, Michael I. Jordan: “Latent Dirichlet Allocation”, *Machine Learning Research*, Vol.3, pp.993-1022, 2003.

[3] 大澤幸生, ネルス E. ベンソン, 谷内田正彦: “KeyGraph : 語の共起グラフの分割・統合によるキーワード抽出”, 電子情報通信学会論文誌, Vol.J82-D-I, No.2, pp.391-400, 1999.

[4] 落合恵理香, 小林一郎: “商品の評価を対象としたレビュー文書の分析”, 言語処理学会 第 18 回年次大会 発表論文集, pp.1176-1179, 2012.

[5] 芹澤翠, 小林一郎: “文書内のトピック数を考慮したトピック追跡の試み”, 言語処理学会 第 18 回年次大会 発表論文集, pp.1196-1199, 2012.

[6] 立川華代, 小林一郎: “文書から取得した制約知識に基づく潜在的トピック抽出”, 言語処理学会 第 18 回年次大会 発表論文集, pp.313-316, 2012.

[7] bayon - a simple and fast clustering tool - Google Project Hosting http://code.google.com/p/bayon/

[8] 田中努, 徳久雅人, 村上仁一, 池原悟: “結合価パターンへの情緒生起情報の付与”, 言語処理学会 第 10 回年次大会 発表論文集, pp.345-348, 2004.

[9] 黒住亜紀子, 村上雄弥, 徳久雅人, 村上仁一, 池原悟: “結合価パターン辞書における情緒表現性のある用言の意味分析”, 電子情報通信学会ソサイエティ大会講演論文集, 基礎・境界, p.168, 2006.

[10] 池原悟, 宮崎正弘, 白井諭, 横尾昭男, 中岩浩巳, 小倉健太郎, 大山芳史, 林良彦: “日本語語彙大系”, 岩波書店, 1997.

[11] Yahoo!LDA, https://github.com/shravanmn/Yahoo_LDA

[12] MeCab: Yet Another Part-of-Speech and Morphological Analyzer
http://mecab.googlecode.com/svn/trunk/mecab