1L2-OS-17a-1

三位一体アプローチによるテキストデータモデリング法の開発 ―宿泊施設の口コミデータを用いた評価推論モデルの構築―

Development of Modeling Approach Using Text Data by Integrating Three Technologies
Construction of Computational Evaluation Model Using Hotel Review Data

野守　耕爾[*1]
Koji Nomori

神津　友武[*1]
Tomotake Kozu

[*1] 有限責任監査法人トーマツ　デロイトアナリティクス※
Deloitte Analytics, Deloitte Touche Tohmatsu LLC

This study proposes a modeling approach using text data based on three technologies: text mining, PLSA and Bayesian Network. The approach enables us not only to understand the current state but also to simulate the changes of the state under different conditions. In this paper, the approach is applied to hotel review data.

1. はじめに

急増する電子化されたテキスト情報とテキストマイニングツールの普及に伴い，テキストデータからいかに有用な知識を抽出するかということが課題となっている．近年ではテキストマイニングの適用事例も増えてきており，コールセンターの対応履歴や顧客満足度調査の自由記述回答，営業日報，Web 上の書き込みなど，様々な分野で適用され経営に活用されている．しかし従来のテキストマイニングは，テキストデータそれ自体の中身の把握をして，改善すべき点やニーズを抽出する際に有効な手段であるが，現状把握に留まっている．

本研究では，テキストデータとその属性データから，条件を変化させたときに，結果がどの程度変化するのかシミュレーション可能にする推論モデルを構築する．

2. 三位一体アプローチによるテキストデータモデリング

本研究では，テキストマイニング，PLSA（確率的潜在意味解析），ベイジアンネットワークという 3 つの手法を統合することによって，テキストデータから現状の結果を把握するだけでなく，条件を変化させたときの結果を推論可能にするモデリングアプローチを提案する．これを図 1 に示す．

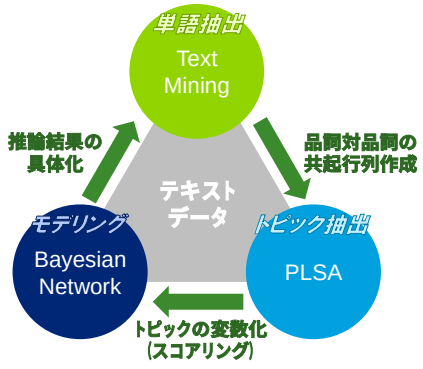

図 1　三位一体アプローチによるテキストデータモデリング法

※本研究の内容は有限責任監査法人トーマツの公式見解を示すものではありません．

連絡先：野守耕爾，有限責任監査法人トーマツ　デロイトアナリティクス，koji.nomori@tohmatsu.co.jp

本アプローチは，①テキストデータにテキストマイニングを適用し単語を抽出する，②抽出した単語で構成される共起行列に PLSA を適用することでテキストデータのトピックを抽出する，③抽出したトピックを変数化し，ベイジアンネットワークを適用することでモデルを構築する，という 3 ステップから構成される．本稿では，宿泊施設の口コミデータを例題に本アプローチの内容について述べ，どのような宿泊条件ではどのようなトピックの口コミがされ，どのようなトピックの口コミがされると満足度にどれほどの影響を与えるのか定量的に推論可能なモデルを構築する．

3. 宿泊施設の口コミデータ

本アプローチの適用例として，旅行情報サイトにおける宿泊施設の口コミデータを用いる．使用データについて表 1 に示す．

表 1　使用データの内容

対象	京都府の「京都駅周辺」「河原町・烏丸・大宮周辺」にある宿泊施設及び口コミの情報
期間	投稿日が2012年5月16日～2013年5月16日
対象宿泊施設数	169件
宿泊施設の取得情報項目	ホテル名，施設タイプ，チェックイン/アウト時間，駅・コンビニまでの徒歩時間，駐車場の有無，温泉・露天風呂・サウナ等の設備の有無，バー・宴会場・カラオケ・屋内プール・禁煙ルーム・製氷機等の施設内容の有無，クリーニング・ルームサービス・マッサージ・チャペル・デイユース等のサービスの有無，貸自転車・囲碁・将棋等の貸し出しレジャーの有無，バストイレ・テレビ・衛星放送・冷蔵庫・ズボンプレッサー等の標準的な部屋設備の有無，有線LAN・無線LAN・PC貸し出し・インターネット接続無料等のインターネット設備の有無，温水洗浄トイレ・ドライヤー・タオル・バスローブ・浴衣・パジャマ・シャンプー・歯ブラシ・髭剃り・綿棒等のアメニティの有無
対象口コミ件数	11,535件　（文章単位で60,958文）
口コミの取得情報	性別，年代，投稿日，項目得点（総合・部屋・風呂・朝食・夕食・接客サービス・清潔感），旅行目的，宿泊価格帯，宿泊部屋タイプ，食事の有無，口コミテキスト

4. テキストマイニングによる単語抽出

第 1 のステップでは，テキストデータにテキストマイニング（形態素解析）を実行することで，各品詞の単語を抽出する．

本研究では，（株）NTT データ数理システムの Text Mining Studio 4.2 を使用しテキストマイニングを実行した．口コミという施設やサービスの評価に関する情報をテキストデータから抽出するため，テキスト内に含まれる名詞と形容詞に着目した．特に形容詞と係り受け関係を持つ名詞，名詞と係り受け関係を持つ形容詞を抽出し，文章単位で出現頻度が 30 件以上の単語に限定したところ，名詞 287 語，形容詞 111 語が抽出された．

5. PLSA によるトピック抽出

第 2 のステップでは，テキストマイニングによって抽出された単語に基づいて PLSA を適用することで，テキストに記述されているトピックを抽出する．

5.1 PLSA(確率的潜在意味解析)

PLSA(Probabilistic Latent Semantic Analysis)は，文章分類に用いるクラスタリング手法として提案され[Hofmann 1999]，文章とそこに出現する単語の間には潜在的な意味クラスがあることを想定したモデルで，文章と単語の共通のトピックとなるような特徴を見つける手法である．PLSA が出力結果において他のクラスタリング手法と異なる点は主に以下の 2 つがある．

① 行と列を同時にクラスタリングする
PLSA では文章 d (行) と単語 w (列) の共起行列を学習データとし，共起行列の行と列にある変数が共通のクラス c に所属する．

② ソフトクラスタリングである
変数が必ず1つのグループに所属するハードクラスタリングと異なり，全ての変数は複数のクラスにまたがって所属し，その所属確率 $P(d|c)$, $P(w|c)$が与えられる．

PLSA ではクラス数をあらかじめ設定する必要があるが，AIC などの情報量基準により最適なクラス数を決定することができる．例えば，クラス数の異なる分析結果それぞれについて AIC を計算し，AIC 最小となるクラス数の結果を採用すればよい．また PLSA は初期値依存性があり，初期値によって結果が異なる．そこで，クラス数を範囲を持たせて複数設定し，初期値を変えてそれぞれのクラス数で PLSA を複数回実行し，その全結果の中で AIC 最小となる結果を採用するといった対応が考えられる．

5.2 提案アプローチにおける PLSA の活用

PLSA は本来，「文章」と「単語」の共起行列に基づき，文章と単語の背後にあるクラスを抽出する手法だが，本アプローチにおける PLSA の活用の仕方は，文章に含まれる単語の「品詞」と「品詞」の背後にあるクラスを抽出する．これは文章を分類することよりも，文章内で記述されているトピックをより明確な形で抽出することが本アプローチでは重要となるためである．例えば評価に関わるトピックを抽出したいときは「名詞」と「形容詞」の単語の共起行列を，行動に関わるトピックを抽出したいときは「名詞」と「動詞」の単語の共起行列を作成し，潜在クラスを抽出する．分析結果では各品詞の単語が共通のクラスに所属するため，そのクラスの意味するトピックを解釈しやすくなる．

5.3 口コミデータの評価トピックの抽出

本研究では，(独)産業技術総合研究所の開発したサービス店舗支援システム APOSTOOL の PLSA プログラムを使用した．

テキストマイニングによって抽出された名詞 287 語と形容詞 111 語の文章単位における共起行列を作成し学習データとした．クラス数を 15 から 25 まで 1 刻みで変化させ，それぞれに対して PLSA を初期値を変え 5 回ずつ実行し，AIC を計算した．その結果，クラス数に対して下に凸のカーブを描き，クラス数 18 の実行結果の一つが AIC 最小となり，この結果を採用することとした．

採用した実行結果について，それぞれのクラス C_k における名詞 N_i と形容詞 A_j をクラスの所属確率 $P(N_i|C_k)$, $P(A_j|C_k)$の高い順に並べ，そのクラスが意味する評価トピックを解釈した．18 個のクラスのうち 3 つのクラスの結果を例に表 2 に示す．C5 のクラスは，部屋の綺麗さに関するトピック，C7 は朝食の美味しさに関するトピック，C14 はスタッフの丁寧さに関するトピックと解釈できる．今回抽出された 18 個のクラスについて同様に意味するトピックを解釈した結果を表 3 に示す．

表 2 PLSA によって抽出されたクラス例

C5				C7				C14			
P(N|C)	名詞	P(A|C)	形容詞	P(N|C)	名詞	P(A|C)	形容詞	P(N|C)	名詞	P(A|C)	形容詞
30%	部屋	29%	綺麗	26%	朝食	59%	美味しい	18%	対応	23%	丁寧
8%	ホテル･宿	25%	清潔	5%	バイキング	6%	良い	11%	フロント	18%	良い
5%	満足	13%	広い	4%	満足	5%	豊富	11%	スタッフ	17%	親切
4%	お風呂	6%	良い	4%	種類	5%	残念	6%	接客	7%	気持ち良い
2%	駅	4%	新しい	3%	パン	3%	大変	5%	ホテル･宿	6%	素晴らしい
2%	利用	4%	快適	3%	料理	3%	十分	4%	部屋	3%	非常

表 3 PLSA によって抽出された 18 個のクラスの解釈

クラスNo.	クラス名	代表的なトピック
1	部屋環境	煙草やエアコンなど部屋の臭い，空調の効き
2	駅近さ	駅やバス停の近さ，観光地･飲食店の近さ
3	値段手頃さ	値段の安さ･手頃さ，値段に対するサービスの評価
4	良さ	居心地，清潔感，対応，アクセスなど様々な良さ
5	部屋綺麗さ	部屋，風呂，館内の清潔感，綺麗さ
6	チェックイン対応	予定より早い遅い到着の対応，荷物の預かり
7	朝食美味しさ	朝食のおかず，パン，バイキングの美味しさ
8	部屋音環境	隣の部屋･廊下･道路の音，部屋の壁の薄さ
9	ホテル旅行良さ	素敵な宿，雰囲気の良い宿，楽しい旅行の思い出
10	サービス嬉しさ	設備･アメニティ･無料サービスの充実，スタッフの心配り
11	値段コスパ満足	値段の高さ，満足の高さ，コストパフォーマンスの高さ
12	部屋気持ちよさ	部屋の行き届いた清掃，気持ち良い滞在，快適性
13	建物綺麗さ	建物の古さや綺麗さ，外観の印象と異なる内観
14	スタッフ丁寧さ	スタッフの対応の丁寧さ，挨拶･笑顔など接客態度
15	場所便利さ	立地の良さ，観光における利便性，周囲の店舗の多さ
16	朝食多さ	朝食のおかず，バイキングの品数，量
17	部屋風呂悪さ	風呂の狭さ，排水，湯の温度，清潔感
18	部屋風呂広さ	部屋･ベッド･風呂･スペースの広さ，快適性

6. トピックの変数化

次のステップでは PLSA で抽出したクラスの示すトピックと，データから得られるテキスト以外の属性情報との関係をモデル化するため，トピックを変数化し，データのレコード単位にそのトピックのスコアを付与する．

これまでも PLSA によって抽出されたクラスを変数として扱い，他の変数間との関係をモデル化する研究事例がある[石垣ら 2011]．そこでは PLSA の共起行列の行と列の一方が元データのレコードとなるように構成されており，得られたクラスに対するレコード単位の変数化処理は，各レコードで最も関係の強いクラス (レコードを条件としたクラスの条件付確率が最も大きいクラス) を 1 つ割り当て，質的変数として扱う方法となっている．

一方本研究では，各レコードを構成する要素 (テキストに含まれる単語) で以って共起行列を構築しており，各レコードに対するそれぞれのクラスとの関係の強さをその構成要素に基づいてスコアリングし，量的変数として扱う手法を提案する．

6.1 トピック変数のスコアリング手法

1 件のレコードにあるテキストデータは複数の文章から構成され，文章によって記述されている話題が異なることがある．本研究では，文章単位にトピックのスコアを計算し，その後レコード単位に各トピックのスコアを集約する方法を提案する．

本手法では，文章 D_h におけるクラス C_k のスコアを $P(D_h|C_k)$ で定義する．そのクラスのトピックを良く表現している文章ほどこの確率は高くなる．PLSA の学習データとなる共起行列が文章に含まれる名詞 (行) と形容詞 (列) で構成されるとき，文章 D について，名詞によって定義される文章を Dn，形容詞によって定義される文章を Da とする．このとき Dn_h と Da_h は定義の仕方が異なるだけでどちらも同じ文章 D_h を意味している．PLSA における文章と品詞とクラスの関係モデルを図 2 に示す．

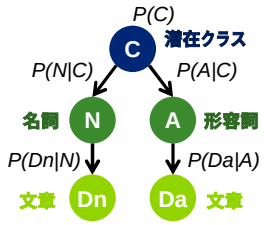

図 2 PLSA における文章と品詞とクラスの関係

$P(D_h|C_k)$を計算するにあたり，$P(Dn_h|C_k)$と $P(Da_h|C_k)$を計算する．これらはそれぞれ式(1),(2)で計算される．単語 w が含まれる文章の数を $n(w)$とすると，$P(Dn|N_i)$と $P(Da|A_j)$はそれぞれ $n(N_i)$ と $n(A_j)$の逆数として計算される．$P(N_i|C)$と $P(A_j|C)$は PLSA の実行結果によって得られる．このとき式(3)が成立し，$P(Dn|C)$と $P(Da|C)$は文章 D において重みは同じといえるので，式(4)により $P(D_h|C_k)$を計算する．

$$P(Dn_h|C_k) = \sum_i P(Dn_h|N_i)P(N_i|C_k) \quad (1)$$

$$P(Da_h|C_k) = \sum_j P(Da_h|A_j)P(A_j|C_k) \quad (2)$$

$$\sum_h P(Dn_h|C_k) = 1, \quad \sum_h P(Da_h|C_k) = 1 \quad (3)$$

$$P(D_h|C_k) = \frac{1}{2}P(Dn_h|C_k) + \frac{1}{2}P(Da_h|C_k) \quad (4)$$

また $P(D_h|C_k)$は文章における総和が 1 となり，元データにおける文章の数が多いほど値は小さくなるが，この値だけではクラスと文章の関係の強さが分かりにくい．そこで事後確率 $P(D_h|C_k)$と事前確率 $P(D_h)$の比となる，$P(D_h|C_k)/P(D_h)$ をもって文章 D_h におけるクラス C_k のスコアとする．この値が 1 を超えるということは，文章 D_h の発生確率はクラス C_k を条件とすることで上昇し，クラス C_k との関係が強いということである．本研究では事前確率は一様分布とし，$P(D_h)$は全文章数の逆数とする．

ここまで文章単位にクラスのスコアを計算したが，これを集約して元データのレコード単位におけるクラスのスコアを決定する．本手法では，文章単位のスコアをレコード単位で見たとき，各クラスのスコアの最大値をそのレコードのクラスのスコアとする．

6.2 口コミの評価の極性判定法

上記の方法で，口コミ 11,535 件に対して 18 個のトピックのスコアを計算し変数化した．しかし，口コミデータに PLSA を適用して抽出されたクラスとは，あくまで評価のトピック（評価視点）であり，その中にはポジティブな意味とネガティブな意味が混在するケースがある．例えば表 2 の C7 の朝食の美味しさに関するクラスの形容詞では，所属確率の高い単語は「美味しい」や「良い」というポジティブな表現だが，「残念」というネガティブな表現も上位語として現れている．そこで，不評文章に関しては計算したクラスのスコアを負数とする．本研究では口コミデータの全文章の極性（好評･不評）を自動で振り分ける方法を以下に提案する．

共起行列に採用した名詞 287 語，形容詞 111 語の口コミ単位における出現有無(0,1)を説明変数に，各口コミにおいてユーザが付与した 6 つの項目得点（「総合」「部屋」「風呂」「朝食」「接客サービス」「清潔感」）を目的変数に 6 つの重回帰モデルを構築した．各単語において，6 つのモデルの偏回帰係数の平均をポジネガポイントとして定義した．例えば，「素晴らしい」は 0.26，「快適」は 0.19，「カビ」は-0.38，「髪の毛」は-0.60 であった．各文章において出現単語のポジネガポイントの合計が 0 未満を不評文章，0 以上を好評文章(中立含む)とした．

判定精度を測定するため，500 件の口コミ文章をランダムサンプリングし，それらを目視で好評文章と不評文章に分け（好評文章 412 件，不評文章 88 件），本手法での判定結果と比較した．再現率を検証したところ，好評文章で約 85%，不評文章で約 70%で，ある程度の精度で振り分けられることが分かった．

従来，評価の極性に関する辞書を構築する手法の研究は，シソーラスの情報やコーパスの共起情報を利用するものなど，様々に取り組まれている．本手法の特長は，主観による極性判別ではなく，評価得点を教師とした客観的な極性判別なので再現性があることと，また定量的な評価尺度と紐づくテキストデータさえあれば，汎用辞書に依存しない，このデータだけに特化した好評語と不評語の辞書を構築できる点にある．

6.3 口コミの評価トピックの極性付変数化

口コミデータに PLSA を適用することで 18 個のトピックが抽出されたが，これに対して，好評文章のスコアを割り当てるポジティブトピック 18 個（変数名 Cp1~18）と，不評文章のスコアを割り当てるネガティブトピック 18 個（変数名 Cn1~18）を設定し，計 36 個のトピックを変数とした．6.1 で示した手順に従い，この文章単位のスコアを口コミ単位で見たとき，36 個のそれぞれのトピック変数のスコアの最大値をその口コミのスコアとして採用した．

7. ベイジアンネットワークによるモデル化

データの中のテキスト情報から抽出されたトピック変数と，テキスト以外の属性情報も変数に採用し，ベイジアンネットワークを適用することで，テキストのトピックと属性情報との確率的関係をモデル化する．これによりどのような条件ではどんなトピックの記述がされるのか，あるいはどんなトピックの記述がされると結果はどの程度影響するのかなど，その関係構造を把握でき，また与えた条件下での確率推論が可能となり，条件を変化させたときの結果の効果を定量的にシミュレーションできる．

これまでもテキストデータからマイニングされた単語情報を変数とし，ベイジアンネットワークによりモデル化する事例はある[野守ら 2010]．しかし単語の出現の有無をそのまま変数としているため，ノードがとても多くモデルが非常に複雑となっている．この場合ベイジアンネットワークのモデルのベースとなる条件付確率表も疎になりやすく，正しい推論ができない可能性が生じてしまう．テキストマイニングとベイジアンネットワークを直接連結させるのではなく，本研究のアプローチのように PLSA を介することで，単語ではなくトピックを変数として扱えるのでモデルがシンプルとなり，結果の解釈もしやすくなる．

7.1 宿泊施設の評価推論モデルの構築

本研究では，宿泊施設の口コミデータを用いて以下の 2 つのモデルを構築した．なお，本研究では，（株）NTT データ数理システムの BAYONET6.1 を使用してベイジアンネットワークのモデルを構築した．

① 評価構造モデル(図 3)
ユーザ属性(性別や年代等)，宿泊内容(宿泊料金や部屋サイズ等)，施設属性(設備やサービス，アメニティ等)といった各種宿泊条件と評価トピックや項目得点との関係をモデル化した．構造条件として，ユーザ属性，宿泊内容，施設属性を親ノード候補に設定した．

② トピック満足モデル(図 4)
評価トピックと項目得点との関係をモデル化した．構造条件として，評価トピックを親ノード候補に設定した．

①評価構造モデルでは，各宿泊条件はどのような評価トピックと関係し，評価得点にどのような影響を与えるのか，その構造を把握し，様々に与えた条件下における確率推論を実行できる．②トピック満足モデルは，どのような評価トピックが各項目得点に影響を与えているのか，その構造を把握し，宿泊施設がどのようなサービスを充実化させると高満足度を得られる確率がどの程度変化するのか推論することができる．

全てのサービス要素を充実させることが最も理想であるが，構築されたモデルの構造と推論結果から，何を優先的に充実させるべきか把握することができる．例えば「総合得点」に着目すると，①評価構造モデルでは，「宿泊部屋サイズ」「パジャマ」「PC 貸出あり」「無線 LAN」という宿泊条件の関連性が高いことが分か

った．②トピック満足モデルでは，好評トピックでは「Cp5:部屋綺麗さ」「Cp7:朝食美味しさ」「Cp9:ホテル旅行良さ」「Cp14:スタッフ丁寧さ」，不評トピックでは「Cn1:部屋環境」「Cn17:部屋風呂悪さ」の関連性が高いことが分かった．

なおベイジアンネットワークで扱う確率変数は全て質的変数となるため，量的変数のカテゴリ化を行った．トピック変数はスコアが 3 超の場合”High”，3 以下の場合”Low”と 2 つのカテゴリを設定した．また各項目得点は 1 点から 5 点までの値を取るが，(1)3 点以下，(2)4 点，(3)5 点，というカテゴリ化ルールとした．

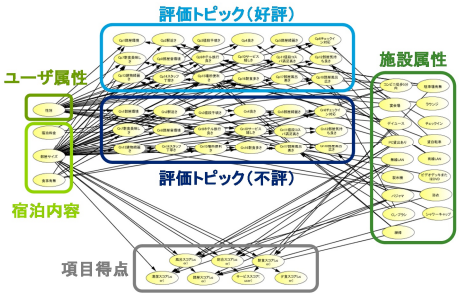

図 3　評価構造モデル

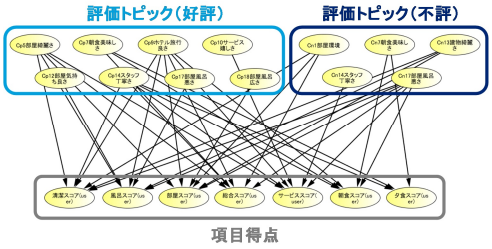

図 4　トピック満足モデル

7.2　モデルを用いた確率推論

構築したモデルを用いて，与えた条件下での確率推論を実行した．その結果例えば，「総合得点」に関係のあった好評トピック「Cp5:部屋綺麗さ」「Cp7:朝食美味しさ」「Cp9:ホテル旅行良さ」「Cp14:スタッフ丁寧さ」と，不評トピック「Cn1:部屋環境」「Cn17:部屋風呂悪さ」を条件とし，「総合得点が 5 点満点」となる確率を推論した結果を図 5 に示す．数多くあるサービスの評価観点の中でも，宿泊客の総合的な満足度に特に寄与するのはこのような観点であり，スタッフが丁寧であることや，部屋が綺麗であることを感じると高満足度を押し上げ，臭いなど部屋の空気が悪いことや，風呂の不具合，不衛生などを感じると高満足度を押し下げることが定量的に把握できる．

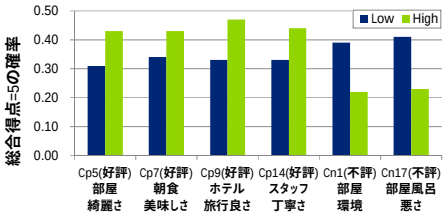

図 5　各トピックに対する総合得点=5 の確率

8.　モデルを用いた推論結果の具体化

最後に，モデリングの結果から得られた有効な変数に焦点を当て，テキストデータの原文を参照したり，再度テキストマイニングを実行して，重要となる具体的なサービス要素を抽出する．

図 5 で総合得点 5 点満点の確率を押し上げる好評トピック「Cp14:スタッフ丁寧さ」と，押し下げる不評トピック「Cn17:部屋風呂悪さ」について，スコアのカテゴリが”High”に割り当てられている文章に限定し（Cp14:4192 件，Cn17：4012 件），そこに含まれる名詞と形容詞との係り受け表現を抽出した．抽出結果の例を表 4 に示す．リフト値は対象文章における頻度の割合を全体文章における頻度の割合で除した値であり，値が大きいほどそのトピックにおいて特徴的な表現であるといえる．

表 4 より，総合満足度を押し上げる好評トピック「Cp14:スタッフ丁寧さ」とは，具体的には，対応が丁寧，親切であることは当然そうだが，笑顔が素敵であることや，嫌な顔をしないこと，挨拶が気持ち良いことといった口コミがされており，宿泊客はスタッフの表情や挨拶をよく見ていることが分かる．また総合満足度を押し下げる「Cn17:部屋風呂悪さ」とは，具体的には，単に部屋や風呂が狭いことだけでなく，排水や流れが悪いことの口コミも多いことが分かる．水圧の高さはホテルの評価ポイントとしてよく知られているが，水圧だけでなく排水がきちんとされることも確認すべきサービス要素であるといえる．

推論モデルから得られた結果を深堀り分析の軸としてとらえ，こうした具体的な口コミの内容を確認して，サービスの質向上に向けた改善施策，投資施策を立案することが重要といえる．

表 4　トピックスコアが”High”の文章に含まれる係り受け表現

「Cp14:スタッフ丁寧さ」の係り受け表現

係り元単語	係り先単語	頻度	リフト値
対応	良い	457	13.2
丁寧	対応	209	14.1
親切	対応	126	14.2
接客	丁寧	80	14.0
フロント	親切	46	13.4
気持ち良い	対応	42	14.2
笑顔	素敵	26	13.5
丁寧	説明	22	10.7
迅速	対応	21	12.2
挨拶	気持ち良い	11	14.5
気遣い	素晴らしい	11	14.5
嫌	顔+ない	10	14.5

「Cn17:部屋風呂悪さ」の係り受け表現

係り元単語	係り先単語	頻度	リフト値
部屋	狭い	222	13.8
お風呂	狭い	153	14.6
気持ち	悪い	45	14.9
狭い	感じ	31	13.5
排水	悪い	26	15.2
トイレ	狭い	22	14.5
浴槽	狭い	14	15.2
流れ	悪い	13	15.2
脱衣所	狭い	12	15.2
冷蔵庫	小さい	12	13.0
駐車場	狭い	11	15.2
テレビ	小さい	9	13.7

9.　まとめ

本研究では，テキストマイニング，PLSA，ベイジアンネットワークという 3 つの手法を統合することで，テキストデータから現状の結果を把握するだけでなく，条件を変化させたときの結果をシミュレーション可能にするモデリングアプローチを提案し，宿泊施設の口コミデータに適用した．本アプローチを用いることで，テキストデータのトピックを抽出し，どのような条件でそのトピックが出現し，またそのトピックが出現した場合はどのような結果となるのか，様々な条件下で定量的な推論が可能となる．

例えば宿泊施設の口コミデータに適用した結果を用いることで，宿泊客はどんな観点の評価軸を持ち，満足度を高めるにはどのようなサービス価値を充実化すべきか把握でき，施策検討のエビデンスとなり得る．また現在提供していると見込まれるサービス価値を条件としたとき，またある改善策や設備投資をしたと仮定したときの，それぞれの条件下における満足度の確率を計算し，その結果から施策の効果を比較することができる．

参考文献

[Hofmann 1999]　Hofmann, T.: Probabilistic latent semantic analysis, Proc. of Uncertainty in Artificial Intelligence, pp. 289-296, 1999.

[石垣ら 2011]　石垣司，竹中毅，本村陽一: 日常購買行動に関する大規模データの融合による顧客行動予測システム：実サービス支援のためのカテゴリマイニング技術，人工知能学会論文誌, Vol.26, No.6, pp.670-681, 2011.

[野守ら 2010]　野守耕爾，北村光司，本村陽一，西田佳史，山中龍宏，小松原明哲: 大規模傷害テキストデータに基づいた製品に対する行動と事故の関係モデルの構築：エビデンスベースド・リスクアセスメントの実現に向けて，人工知能学会論文誌, Vol.25, No.5, pp.602-612, 2010.