

旅行情報サイトにおけるホテル口コミデータの テキストマイニング

Text Mining of Review's Data of Hotel in Internet Travel Advisory Service

久保田 貴 文*
Takafumi KUBOTA

キーワード： 口コミデータ、テキストマイニング、対応分析

1. はじめに

近年のインターネットの普及により、ホテルの予約形態が変わりつつある。従来までの旅行代理店経由や電話による予約に比べて、インターネットから予約する機会が増加している。

加えて、「国内旅行の宿泊予約」についてアンケート調査（DIMSDRIVE、2008）によると、ホテルを予約する際には、価格やアクセスのような量的変数の他にも、いわゆる口コミといわれる「お客様の声」、すなわち質的変数を重視することも示されている。さらに、予約をする際には、ホテルのHPなどで直接予約をするよりも、旅行情報サイトを利用する方が多い割合を占めている（DIMSDRIVE、2008）。

旅行情報サイトのデータを用いたデータ解析の先行研究として、高野 他（2009）では、楽天トラベル（楽天、2015）のサイト上の口コミデータを用いてテキストマイニングにより因果関係に着目して評価表現を抽出しており、また、石橋（2012）では、伊豆地域のホテル・旅館を対象として、じゃらんnet（リクルート、2015）のサイト上の口コミデータを用いてテキストマイニングによりブランド・評判の改善等に貢献することを目的としている。これらの研究では、旅行情報サイトの口コミデータを実際の利用者のホテルを選ぶ観点からは解析されていない。

そこで、本論文では楽天トラベルの口コミデータを用いて、第一にテキストマイニングを行い、日本におけるホテルの口コミの傾向を掴んだ。第二に頻出する単語についてデータを再構成し、対応分析により単語間・ホテルグループ間の対応関係を調べ、クラスター分析によりホテルグループをいくつかの集団にわけて考察した。最後に、その解析結果から代表的な3つのホテルをピックアップしさらにテキストマイニングを進めた。

まず、2節にて本論文で利用したデータの形式について述べる。さらに、3節にてテキストマイニングおよび多変量解析の結果を示す。最後に4節で、まとめと今後の展望について述べる。

* 多摩大学経営情報学部 School of Management and Information Sciences, Tama University

2. データについて

データ 1

「国内旅行の宿泊予約」についてアンケート調査（DIMSDRIVE、2008）より、最もよく使われている楽天トラベル（楽天、2015）の口コミデータについて、掲載されている全てのホテルグループに対して、おすすめ順が高い順（2015年7月29日時点）に上位3軒のホテルについて、最新（同時点）の口コミを抽出した（図1）。このうち、口コミのテキストデータを対象にして、解析を行った。

	A	B	C	D	E	
1	ID	グルー	ホテルグループ	ホテル略称	ホテル名	口コミ
2	投稿者さあ01	アー	ークホテルグループ	アー	ークホテル大阪心斎橋	地下鉄の駅から至便であり、この時期の雨に降られる心配なし。チェックアウト時間が遅めだったので、行動プランの自由度が上げられ、朝食が充実しており、おいしくいいます。
3	投稿者さあ01	アー	ークホテルグループ	アー	ークホテルロイヤル福岡天神	また利用させて頂きたいです。
4	投稿者さあ01	アー	ークホテルグループ	アー	ークホテル広島駅南（ルート）	立地は駅から少し遠かったですが、部屋、風呂、朝食は満足のいくも
5	投稿者さあ02	ア	ートホテルズ	ア	ートホテルズ大森	久しぶりに宿泊しました。たまたまだと思いますが部屋が13階のリニ
6	投稿者さあ02	ア	ートホテルズ	ア	ートホテルズ浜松町	部屋は広めでゆったりしています。朝食の種類も豊富です。お掃除が日によって行き届いていないことがありました。
7	投稿者さあ02	ア	ートホテルズ	ア	ートホテルズ札幌	札幌の宿泊として、平成10年頃から利用させていただいている者です。満足しております。また、今回は3連泊でしたから、毎朝朝食を選べる
8	まこちゃんあ03	R & B	ホテルチェーン（ワシントン）	R & B	R & Bホテル熊谷駅前	駅前の好立地で大変良かったです
9	F-TOSHIあ03	R & B	ホテルチェーン（ワシントン）	R & B	R & Bホテル札幌北3西2	いつもと変わらずに 使わせてもらいありがとうございます
10	シータハスあ03	R & B	ホテルチェーン（ワシントン）	R & B	R & Bホテル博多駅前	近隣ですが、騒音の無くてゆっくりできて疲れも取れました。

図1 楽天トラベルから収集した口コミデータの一部

データ 2

データ1を用いてテキストマイニングで解析した結果のうち、頻度が高い20の単語を含むかどうかを、それぞれのホテルグループで集計し、ホテルグループ・単語行列として再編成した（図2）。

	row.names	良い	部屋	利用	朝食	美味しい	対応	立地	宿泊	
1	アー	0	1	2	2	0	0	1	0	
2	ア	0	3	2	3	0	1	1	2	
3	R & B	0	0	1	1	0	0	1	0	
4	ノ	0	1	1	0	0	1	0	0	
5	イ	0	1	0	0	0	0	0	0	
6	ザ サイプレス	0	1	1	1	1	0	0	0	
7	ホテル・ア	0	0	1	1	0	1	0	0	
8	古湯温泉	0	1	0	0	0	0	0	0	

図2 ホテルグループ・単語行列の一部

データ 3

データ2の解析結果から判断し、「ドリーミン」、「アパホテル」および「東横イン」のホテルグループに対して、おすすめ順が高い順（2015年9月28日時点）に上位10軒のホテルについて、最新（同時点）の20件の口コミを抽出した（図3）。

3. 解析結果

本論文では、テキストマイニングによる解析のために、データ1およびデータ3については、Text Mining Studio (NTT データ数理システム、2015) を使い、データ2については、R、MeCab および RMeCab を使った。はじめに、データ1を用いて、頻度分析、クラスター分析、およびネットワーク分析を行った。図4に頻度分析の結果を示す。

	A	B	C	D	E	F
1	id	group	hotelNumber	hotelName	text	
2	1	DU	1	加賀の湧泉 ドーミーイン	初めての利用でした。後から考えれば駅前	
3	2	DU	1	加賀の湧泉 ドーミーイン	台風の影響で急遽宿泊。駅に近く便利で	
4	3	DU	1	加賀の湧泉 ドーミーイン	朝食について、京都と同程度を期待してい	
5	4	DU	1	加賀の湧泉 ドーミーイン	出張で駅近が良く、温泉と食事が良さそう	
6	5	DU	1	加賀の湧泉 ドーミーイン	9月10日に宿泊しました。お風呂が故障し	
7	6	DU	1	加賀の湧泉 ドーミーイン	駅近の立地は観光の拠点として便利。少し	
8	7	DU	1	加賀の湧泉 ドーミーイン	立地は良く、浴場は綺麗で、フロントの印象が	

図3 楽天トラベルから収集した「ドーミーイン」、「アパホテル」および「東横イン」の口コミデータの一部

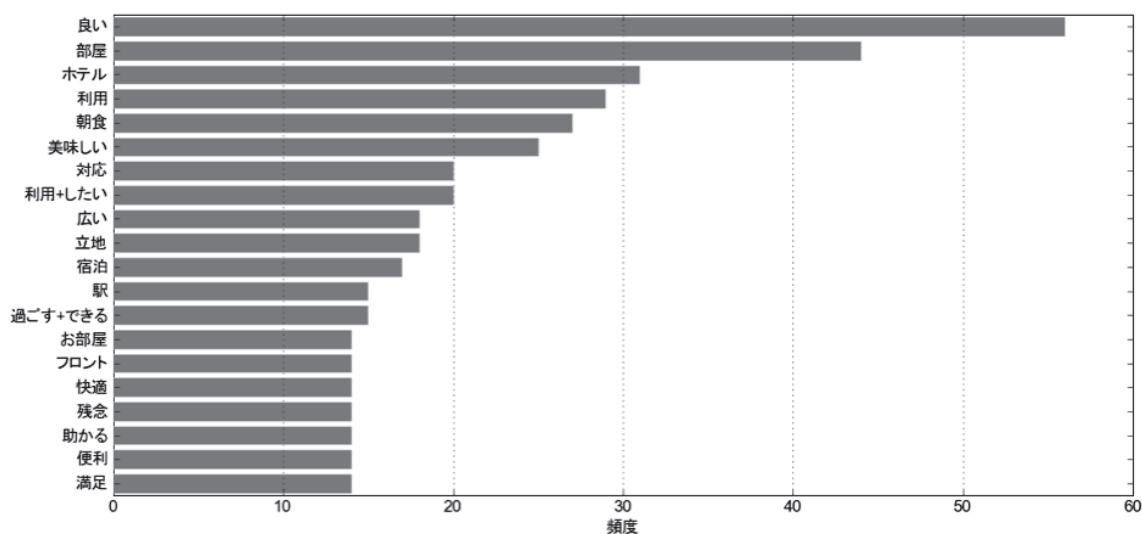


図4 データ1の頻度分析の結果

これより、良い、部屋、ホテル、利用、朝食などが頻繁に口コミで述べられていることがわかった。図5にクラスター分析の結果を円グラフでクラスター毎の割合を示す。これより、ここでも同じように、朝食、部屋、利用、ホテルなどのグループの他に、対応というグループもあることが分かった。また、ネットワーク図(図6)より、良い、部屋、ホテルは様々な言葉と繋がっている一方で、利用は繋がりが薄いことも分かった。

次に、データ2を用いて対応分析を行った。図7に対応分析の結果をバイプロットにて描く。この結果から、横軸は、右に行くほどフロント、対応、立地などビジネス利用、左に行くほど過ごす、美味しい、快適などプライベート利用と解釈し、ビジネス・プライベートの軸と判断できる。また、縦軸は、上に行くほど、便利、立地、駅など立地の項目、下に行くほど、満足、

フロント、対応などサービスの項目であると解釈し、立地・サービス軸と判断できる。この中から、いわゆるビジネスホテルとされるホテルグループのうち、横軸の右に大きい「東横イン」、縦軸の下に大きい「ドーミーイン」、およびどちらも中心にちかい「アパホテル」のグループについてさらに掘り下げて詳しく解析する。(以降、図中には順に、TY、DU、および AP と略す。)

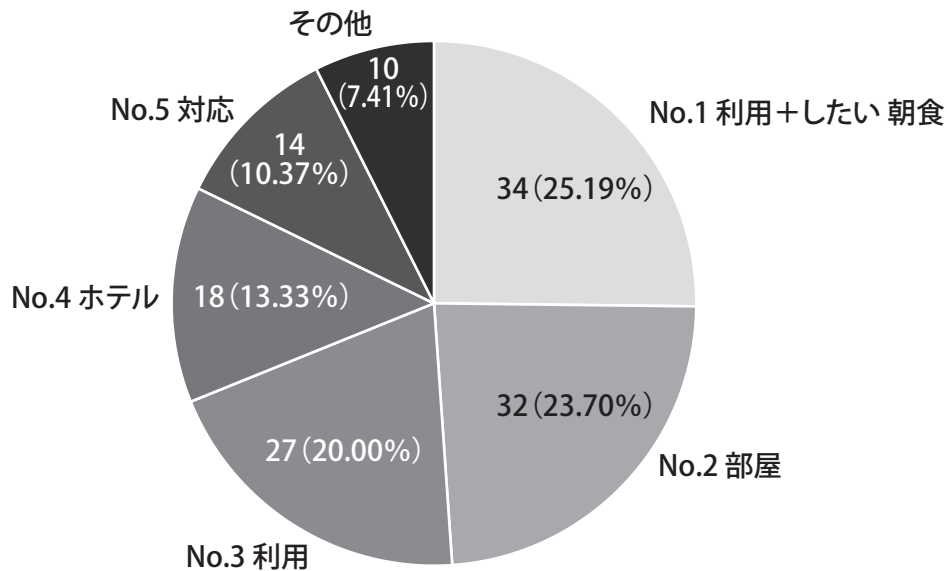


図5 データ1のクラスター分析の結果

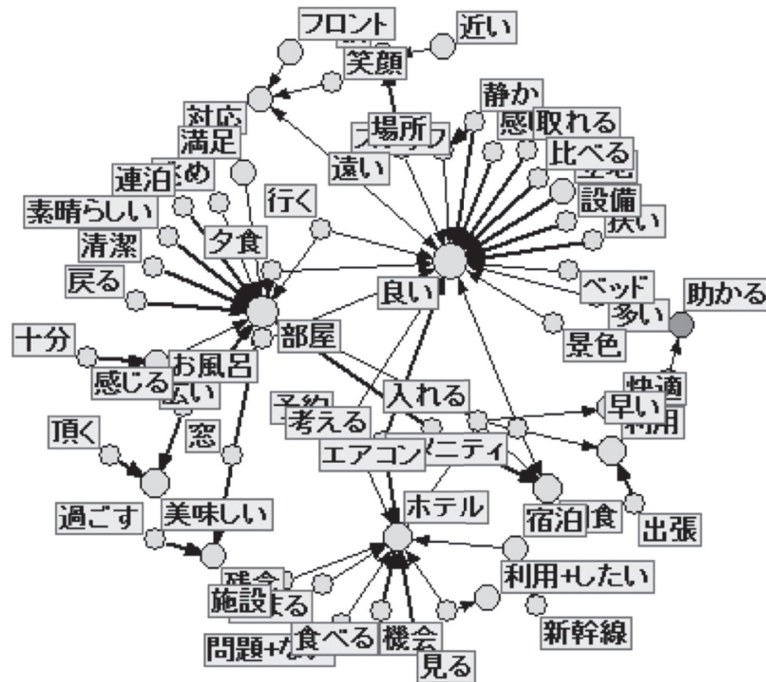


図6 データ1のネットワーク分析の結果

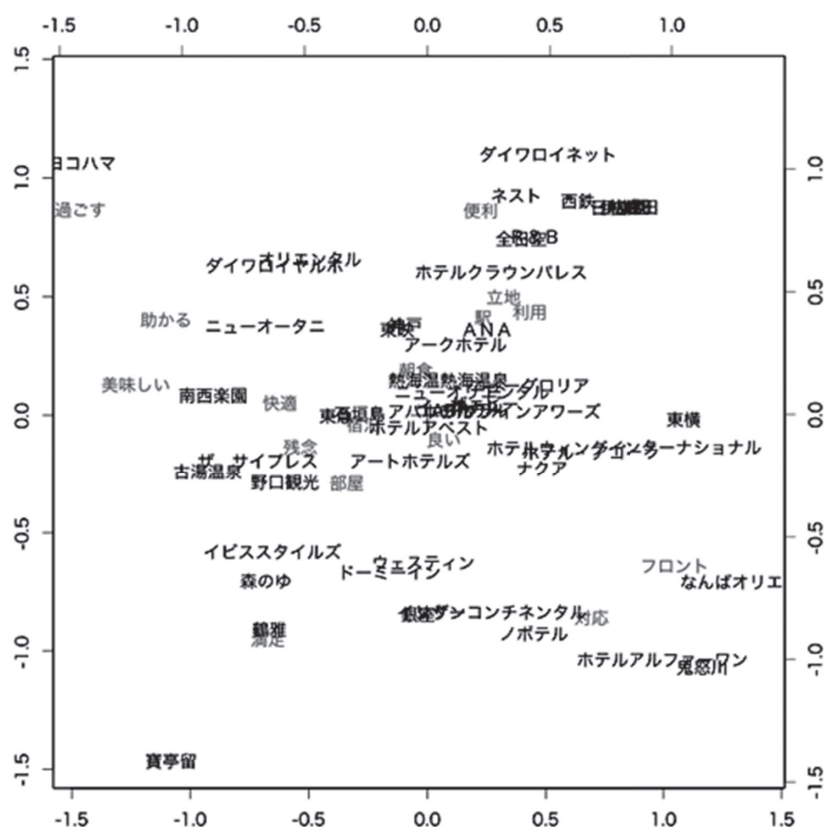


図7 データ2の対応分析の結果（バイプロット）

図8にデータ2を用いたクラスター分析の結果のうちデンドログラムを示す。上側がプライベートで利用が想定されるような温泉などのホテルのグループ、下側がビジネスホテルと判断できるようなグループに分かれている。詳しく解析する3つのホテルは、前者から「ドーマーイン」、後者から、「東横イン」と「アパホテル」である。

最後にデータ3を用いて、テキストマイニングで詳しく解析を行う。3つのホテルグループを対象にしているため、それぞれのグループ毎の頻度分析の結果（図9）、評判分析の結果（図10）を描く。また、ネットワーク分析の結果を図11に示す。図9より、データ1の結果（図4）と大きく変わらないが、東横インには、大浴場やお風呂の口コミがほとんどないことがわかる。また、図10より、アパホテルには部屋に対するネガティブな口コミが、ドーマーインには食事やお風呂に対するポジティブな口コミが、さらに東横インには立地や駅のポジティブな口コミが多いことが分かる。最後に図11よりそれぞれのグループ関連する口コミのワードについて確認された。

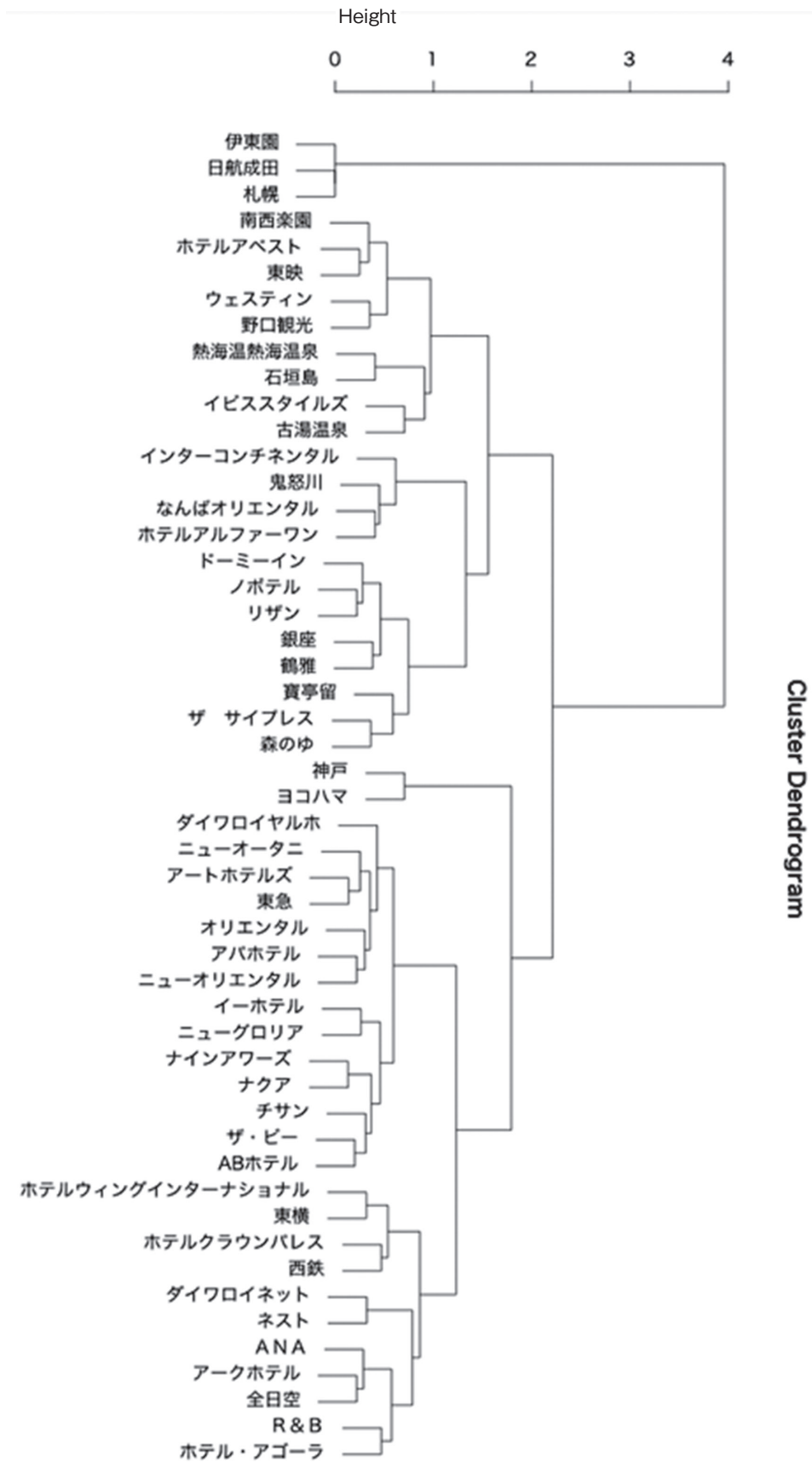


図8 データ2のクラスター分析の結果（デンドログラム）

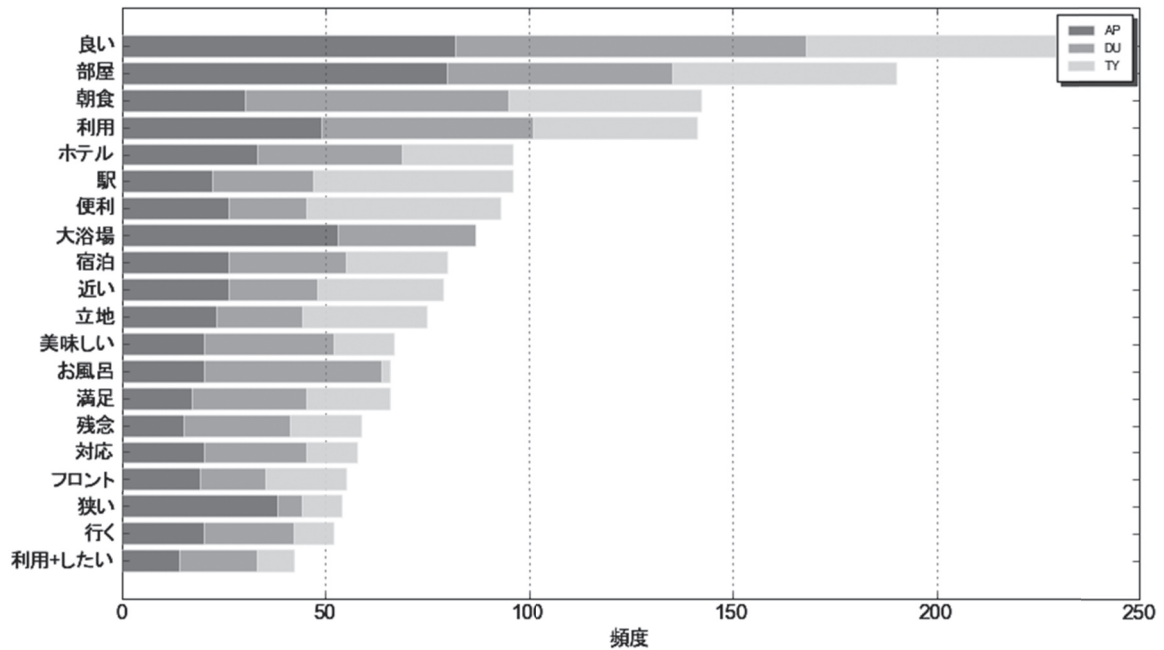


図9 データ3のホテルグループ毎の頻度分析の結果

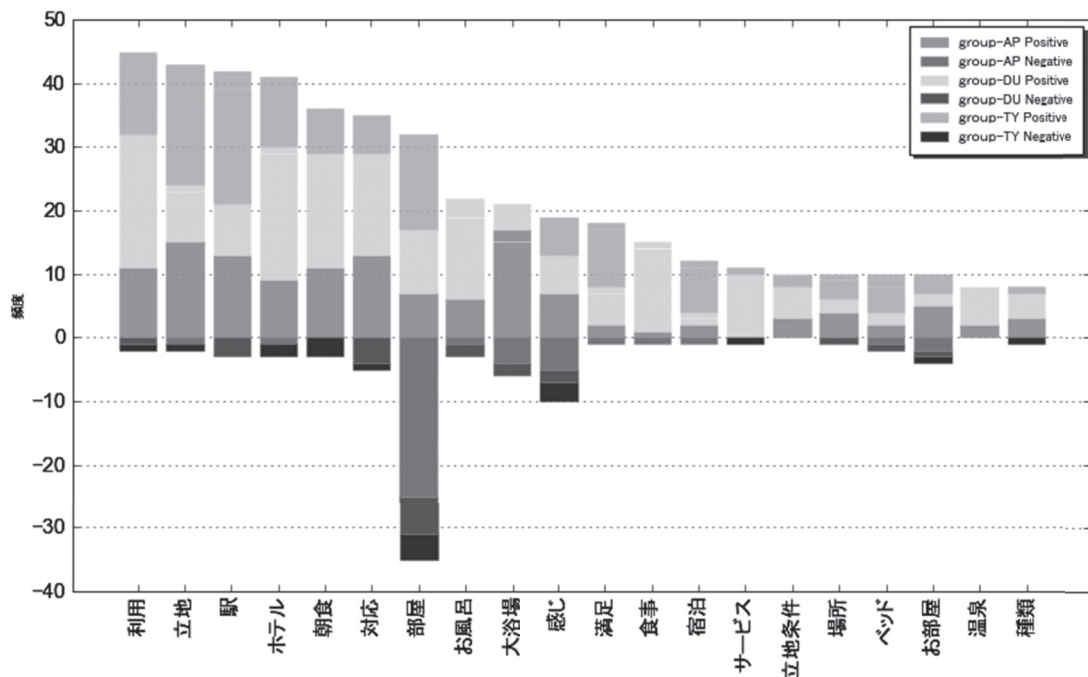


図10 データ3のホテルグループ毎の評判分析の結果

