確率的潜在意味解析を用いた 飲食店・特徴語同時分類結果の飲食店推薦システムへの応用

椎田 太輝[*1] 手塚 太郎[*2] 木村 文則[*3] 前田 亮[*2]

[*1] 立命館大学大学院理工学研究科 〒525-8577 滋賀県草津市野路東 1-1-1
[*2,3] 立命館大学情報理工学部 〒525-8577 滋賀県草津市野路東 1-1-1

E-mail: [*1]cm005055@ed.ritsumei.ac.jp, [*2]{tezuka, amaeda}@media.ritsumei.ac.jp, [*3] fkimura@is.ritsumei.ac.jp,

あらまし ウェブ上の飲食店レビューサイトでは, ユーザに提示される飲食店ランキングにレビュー文の内容が十分に反映されていない. そこで, 本論文では飲食店レビューサイトから収集した飲食店単位のレビュー文から飲食店・特徴語の特徴ベクトルを作成し, それに対して確率的潜在意味解析を適用する. それより取得した飲食店・特徴語の同時分類結果を飲食店の評価・推薦へと応用するシステムの提案を行う.

キーワード 情報推薦, PLSI, レビューサイト, 分類, 嗜好性

1. はじめに

近年, レビューサイトやウェブログのようにインターネット上で社会的なネットワークを構築するサービスの提供・利用が著しい. レビューサイトとは人物・組織・商品・サービスなどの物事に関する評判や噂を扱うウェブサイトの総称である. その中でも飲食店を対象としたレビューサイトはユーザも多く, また“HOT PEPPER”, “ぐるなび”, “食べログ”など数多くの人気飲食店レビューサイト兼検索サイトが存在する. 最近ではそれらのサイト内のコンテンツも充実しており, レビューの評価方法やその提示方法にも様々な工夫がされている.

上記の一般的な飲食店検索サイト兼レビューサイトでは, 飲食店の分類を行い易くするために各飲食店に住所・飲食店カテゴリ・予算・用途などの統計的に処理ができる項目が予め付与されている. そして, ユーザが飲食店を検索する際に選択した検索項目に対し絞り込みを行い, 検索結果として条件に適合する飲食店一覧の提供を行っている. さらに, 各飲食店に対して採点形式により評価された点数やユーザのアクセス数, ユーザからの支持票などを相対的に考慮し, 検索結果の飲食店一覧に対し順位付けを行うことでユーザに対する提示方法を工夫している.

しかし, 上記のような方法の場合, 各飲食店に投稿されたレビュー文の内容は飲食店の順位付けに考慮されることがないため, 飲食店が正しく評価された結果を提示しているとは言い切れない. さらに, 各飲食店の被レビュー数が増加するにつれて, その膨大な量のレビューはユーザ一人あたりの可読量を超えてしまい, 読まれない可能性があるため結果的に無駄になりかねない. つまり, 各飲食店に対するユーザからの評価が飲食店評価に十分に反映されていないことになる.

そこで, 飲食店レビュー文の内容を飲食店評価に反映させるために, 本研究では各飲食店に投稿されたレビュー文に対して確率的潜在意味解（PLSI: Probabilistic Latent Semantic Indexing）[1]を適用することで飲食店の分類を行う. このように, 事前に雰囲気の類似した飲食店を潜在カテゴリに分類することで, レビュー文を飲食店評価に反映させることができる. また, PLSI により飲食店と同時に分類された各潜在カテゴリ中の飲食店の特徴語を飲食店推薦システムに応用する方法について考察を行う.

2. 関連研究

石垣ら[2]は, 流通量販店での顧客行動に基づく大規模履歴データと顧客アンケートデータに対し PLSI を適用することで, 顧客のライフスタイルに基づく顧客と商品の分類を行い, 状況依存性を取り組んだ顧客行動のモデル化を行っている. さらに石垣ら[3]は PLSI を拡張し, 潜在カテゴリを二重に仮定したモデルを使用した顧客と商品の分類を行っている. 本研究では上記 2 つの論文を参考に飲食店単位のレビュー文に適用する PLSI モデルを作成した.

3. 提案手法

本論文で提案する飲食店推薦システムの概要を図 1 に示す. まず, 推薦を行う前に飲食店レビューサイトから飲食店単位でレビュー文を収集し, 特徴ベクトルを作成する. それに対して PLSI を適用する. このとき, 情報量基準を用いて予め潜在カテゴリ数を指定する. その後, 飲食店と特徴語を潜在カテゴリへと同時分類を行う. 本論文では, PLSI を用いて分類された上記 3 集合間（飲食店, 潜在カテゴリ, 特徴語）の関係を推薦システムに応用する方法を提案する. ユーザは飲食店に関するキーワードを特徴語の中から選択すると, それに一致する飲食店一覧がユーザに提供される.

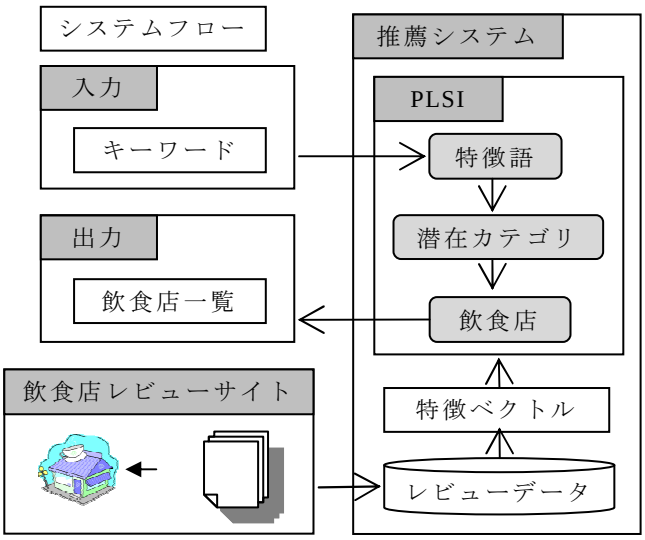
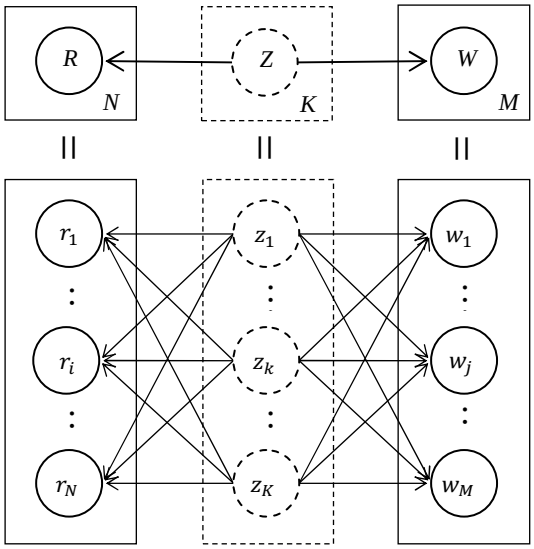


図 1 システムフロー図



(飲食店集合) (潜在カテゴリ集合)(特徴語集合)

図 2 PLSI モデル

4. PLSI アルゴリズム

PLSI とは自然言語処理の分野で文書と単語の特徴ベクトルから文書の意味を推定するために提案された次元縮約法かつ潜在クラス分析の一種である.

本研究においてグラフィカルモデルに基づき作成した PLSI モデルを図 2 に示す. ここでは, N 件の飲食店集合$r \in R = \{r_1, \cdots, r_N\}$とレビュー文書中に出現する M 語の特徴語集合 $w \in W = \{w_1, \cdots, w_M\}$, K 個の潜在カテゴリ集合$z \in Z = \{z_1, \cdots, z_K\}$を用意する. これらの関係は(1)式と表現できる. なお, (1)式は z で周辺化すると(2)式へと変形できる.

飲食店 r に出現する特徴語 w の出現回数を$n(r,w)$と表記すると, 本モデルの対数尤度 L は(2)式を用いて(3)式と表現できる. この対数尤度 L を最大化するパラメータを EM アルゴリズムによる反復計算を用いて推定を行う. このとき推定するパラメータ数は, $N \times K$ 個の $p(r|z)$, $M \times K$ 個の$p(w|z)$, K 個の$p(z)$となる.

E-step では, 上記の各パラメータに対して初期値を乱数で(4)式に与え, 対数尤度 L の期待値$p(z|r,w)$を算出する. その後, M-Step では対数尤度 L の期待値 $p(z|r,w)$を最大化するパラメータ($p(r|z)$, $p(w|z)$, $p(z)$)を求める. なお, このとき対数尤度 L の期待値$p(z|r,w)$を最大化するパラメータはラグランジュの未定乗数法より取得した(5)(6)(7)式となる.

二回目以降の反復計算では, 一回目の M-Step で求めたパラメータを用いて対数尤度 L の期待値を算出する. この反復を対数尤度 L が収束するまで実行し, 対数尤度 L が最大値を記録したときの各パラメータを飲食店と特徴語の分類に用いる.

なお, 予め指定する必要がある潜在カテゴリ数は情報量基準により決定することが可能である.

$$p(r,w,z) = p(r|z)p(z)p(w|z) \quad (1)$$

$$p(r,w) = \sum_z p(r|z)p(z)p(w|z) \quad (2)$$

$$L = \sum_{r,w} n(r,w) \log\{\sum_z p(r|z)p(z)p(w|z)\} \quad (3)$$

$$p(z|r,w) = \frac{p(r|z)p(z)p(w|z)}{\sum_z p(r|z)p(z)p(w|z)} \quad (4)$$

$$p(r|z) = \frac{\sum_w n(r,w)p(z|r,w)}{\sum_{r,w} n(r,w)p(z|r,w)} \quad (5)$$

$$p(w|z) = \frac{\sum_r n(r,w)p(z|r,w)}{\sum_{r,w} n(r,w)p(z|r,w)} \quad (6)$$

$$p(z) = \frac{\sum_{r,w} n(r,w)p(z|r,w)}{\sum_{r,w,z} n(r,w)p(z|r,w)} \quad (7)$$

表 1 不要語処理前と後における各特徴語数の比較

特徴語	処理前	処理後
名詞	28,760 語(76.7%)	16,151 語
動詞	5,752 語(15.3%)	
副詞	1,964 語(5.2%)	
形容詞	853 語(2.3%)	644 語
感動詞	181 語(0.5%)	
合計	37,510 語	16,795 語

表 2 作成した特徴ベクトル情報

特徴ベクトルの大きさ ($N \times M$)	1,200×17,155
総出現頻度	2,136,430
充填率	4.39%

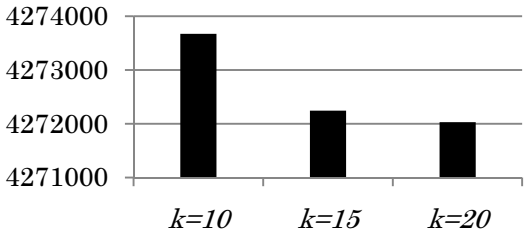


図 3 各潜在カテゴリ数のおける BIC の遷移

5. 飲食店・特徴語の同時分類実験

5.1. 実験データ

本実験で扱う実験データは，レストランガイド「食べログ[4]」内のレビューデータを対象とする．取得した実験データは，京都府にある飲食店のうち，レビューの被投稿件数が多い上位 1,200 件を対象とした．すなわち，各飲食店に投稿された全レビュー文を一単位として計 1,200 件分のデータを用意した．各飲食店に振り分けられたレビューの総文字数は 3,000 語から 230,000 語と大きな差ができたが，補正は行わなかった．そして，それらのレビュー文をまとめて形態素解析器にかけ，飲食店と特徴語の特徴ベクトルを作成した．この特徴ベクトルに対し PLSI を適用することで飲食店と特徴語の同時分類実験を行った．

なお，分類の精度を向上させるために特徴語として 5 つの品詞（名詞・動詞・副詞・形容詞・感動詞）のうち，飲食店の特徴を表しやすいと考えられる名詞と形容詞のみを用いた．さらに，その中から自身で作成した不要語リスト内の特徴語を除いた．それと文書頻度が 2 以下かつ 8 割以上の飲食店に出現する特徴語に対しても除去を行った．上記の方法を用いて形態素解析を行う段階で索引内の不要語を予め除去した．この処理を行うことで不要語と考えられる特徴語 20,715 語を除去した（表 1）． その結果取得した特徴ベクトル情報を表 2 に示す．

5.2. 潜在カテゴリ数の決定

PLSI による分類を行うためには，予め潜在カテゴリの数を指定する必要がある．その潜在カテゴリ数は情報量基準を用いて決定することができる．各潜在カテゴリ数で初期値を変更し，複数回 PLSI を実行して BIC（Bayesian Information Criterion）による情報量基準の変化を調べ，平均的に最適値を示したものを潜在カテゴリ数とする．今回の実験では潜在カテゴリ k=10, 15, 20 の 3 つの場合に対して，初期値を変更して PLSI を各 10 回づつ実行し(8)式に対して平均的に最小値をとる値を潜在カテゴリ数とした．その結果 k=20 が潜在カテゴリとして最適であると判断した（図 3）．

$$BIC = -2L + m\log(N) \quad （m：パラメータ数） \qquad (8)$$

表 3 各潜在カテゴリ内の飲食店数の分布

K	1	2	3	4	5	6	7	8	9	10
店数	42	30	42	80	105	71	51	55	37	52
K	11	12	13	14	15	16	17	18	19	20
店数	57	116	122	38	40	40	48	47	54	66

表 4 各潜在カテゴリ内の飲食店ジャンル分布

K	飲食店ジャンル（件数）		
	1 位	2 位	3 位
k=1(42)	イタリアン(16)	寿司(11)	その他
k=2(30)	ラーメン(11)	その他	その他
k=3(42)	うどん(17)	お好み焼き(14)	その他
k=4(80)	洋菓子(78)	その他	その他
k=5(105)	居酒屋(70)	フライ(15)	その他
k=6(71)	京料理(65)	その他	その他
k=7(51)	ケーキ(51)		
k=8(55)	京料理(53)	その他	その他
k=9(37)	甘味処(35)	その他	その他
k=10(52)	蕎麦(31)	鰻(8)	その他
k=11(57)	焼肉(40)	韓国料理(10)	その他
k=12(116)	カフェ(80)	喫茶店(15)	その他
k=13(122)	イタリアン(60)	フレンチ(50)	その他
k=14(38)	ラーメン(30)	坦々麺(5)	その他
k=15(40)	ラーメン(20)	タイ料理(10)	その他
k=16(40)	和菓子(22)	甘味処(15)	その他
k=17(48)	中華料理(41)	ラーメン(7)	その他
k=18(47)	インド料理(11)	カレー(10)	その他
k=19(54)	ラーメン(54)		
k=20(66)	洋食(30)	ハンバーグ(22)	その他

表 5 潜在カテゴリ（k=1）の特徴語（名詞）上位 20 語

順位	特徴語	$p(w\|z)$	順位	特徴語	$p(w\|z)$
1	寿司	0.031	11	釜	0.003
2	ピザ	0.026	12	ネタ	0.002
3	鯖	0.023	13	ナポリ	0.002
4	和久	0.006	14	名物	0.002
5	酢	0.005	15	巻き	0.002
6	伊勢丹	0.005	16	水牛	0.002
7	マルゲリータ	0.004	17	シンプル	0.001
8	昆布	0.003	18	菱	0.001
9	持ち帰り	0.003	19	デパート	0.001
10	飯	0.003	20	稲荷	0.001

表 6 潜在カテゴリ（k=1）の特徴語（形容詞）上位 20 語

順位	特徴語	$p(w\|z)$	順位	特徴語	$p(w\|z)$
1	分厚い	0.00068	11	香ばしい	0.00006
2	浅い	0.00021	12	気安い	0.00006
3	この上ない	0.00012	13	喜ばしい	0.00005
4	勿体ない	0.00011	14	待ち遠しい	0.00005
5	宜しい	0.00010	15	おこがましい	0.00005
6	偉い	0.00010	16	久しい	0.00004
7	名高い	0.00009	17	物珍しい	0.00004
8	芳しい	0.00008	18	ほの暗い	0.00004
9	色濃い	0.00008	19	しがない	0.00004
10	奥ゆかしい	0.00007	20	華々しい	0.00002

5.3. 分類結果と考察

5.3.1. 飲食店の分類結果と考察

各飲食店の被レビュー件数を一単位とした合計1,200件の飲食店を20個の潜在カテゴリに分類した結果を表3に示す．なお，実験データ収集先である食べログ内で各飲食店に付与されている飲食店ジャンル・予算・用途を用いて各潜在カテゴリ内の飲食店の特徴の確認を行った．

表4は各潜在カテゴリ内の飲食店ジャンルの分布である．表4より，潜在カテゴリ（k=1,3,10,13,15）以外では飲食店のジャンルごと明確に分類されていることが確認できる．しかし，潜在カテゴリ（k=1,3,10,13,15）内の混在するジャンルの飲食店は雰囲気などの共通点があることが読み取れる．このように潜在カテゴリ内に飲食店ジャンルが混在する場合は，PLSIより取得した$p(w|z)$を参照することで飲食店の特徴を把握することができると考えられる．

図4は各潜在カテゴリ内の予算の分布を示す．潜在カテゴリ（k=1,5,6,8,11,13）は比較的高い価格帯であることが読み取れる．一方，潜在カテゴリ（k=7,14,19）は価格帯が特に低いことが確認できる．

図5は潜在カテゴリ内の用途別の分布である．全体的に友達・同僚での用途は均一に分布していることが読み取れる．一方で，デートとひとりの用途は相反する関係にあることが確認できる．これより，両者（デート・ひとり）間の潜在カテゴリ内の特徴語には，これを特徴付ける単語が含まれていることが期待できる．

5.3.2. 特徴語の分類結果と考察

表5と表6は潜在カテゴリ（k=1）における特徴語（名詞・形容詞）の上位20語である．

表5ではイタリアンと寿司に関連する特徴語が占めていることが確認できるが，両者の飲食店ジャンルに共通する名詞を確認することは比較的難しいことが読み取れる．

表6では特定の飲食店ジャンルを特徴付ける形容詞は確認しづらい．しかし，これは各潜在カテゴリ内の飲食店を相対的に修飾する特徴的な単語であることを示していると考えることができる．つまり，潜在カテゴリ中に様々な飲食店ジャンルが混在する場合には飲食店の雰囲気が類似していることを意味する．このことは他の潜在カテゴリ内の飲食店に対しても同様の傾向が確認できた．これより，潜在カテゴリ数をより増やすことでより明確な特徴をもつ飲食店集合へと細かく分類することができると考えられる．

実験結果より，PLSIを用いることで雰囲気が類似する飲食店集合を潜在カテゴリ毎に分類可能であることがわかった．それと同時に各飲食店を特徴付ける特徴語の抽出にもPLSIは適切であることが確認できた．

6. おわりに

本論文では飲食店の被レビュー文を一単位として飲食店計1,200件分の実験データから特徴ベクトルを作成し，そのデータに対してPLSIを適用することで飲食店を潜在カテゴリへと分類した場合の実験結果についての考察を行った．

今後の課題はPLSIより取得した分類結果を飲食店推薦システムに適切に応用する方法を考えることである．今回の実験で取得した特徴語の中には飲食店検索時のユーザの気分や状況，または飲食店に求める雰囲気として利用可能な特徴語が含まれているため，これらを飲食店推薦システムに取り入れていくことが考えられる．

また，レビューには本文だけではなく，レビュアの性別・居住地・年齢や訪問日などのユーザ属性や飲食店の立地条件や最寄り駅などの飲食店属性，また5点満点の採点形式の評価なども含まれているため，それらも構築するシステムの中に取り入れることで，ユーザに対して適切な飲食店を推薦できると考えられる．

参　考　文　献

[1] T. Hofmann, “Probabilistic Latent Semantic Indexing” Proc. the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval, 1999

[2] 石垣司，竹中毅，本村陽一，“確率的潜在意味解析を用いた大規模ID-POSと顧客アンケートの統合利用による顧客-商品の同時カテゴリ分類”，電子情報通信学会技術研究報告 vol.109(461)，pp.425-430，2010

[3] 石垣司，竹中毅，本村陽一，“2重潜在クラスモデルとベイジアンネットワークを結合した小売サービスにおける顧客購買行動モデリング”，情報論的学習理論と機械学習研究会(IBISML)，pp.167-173，2010

[4] “食べログ.com”, http://tabelog.com/

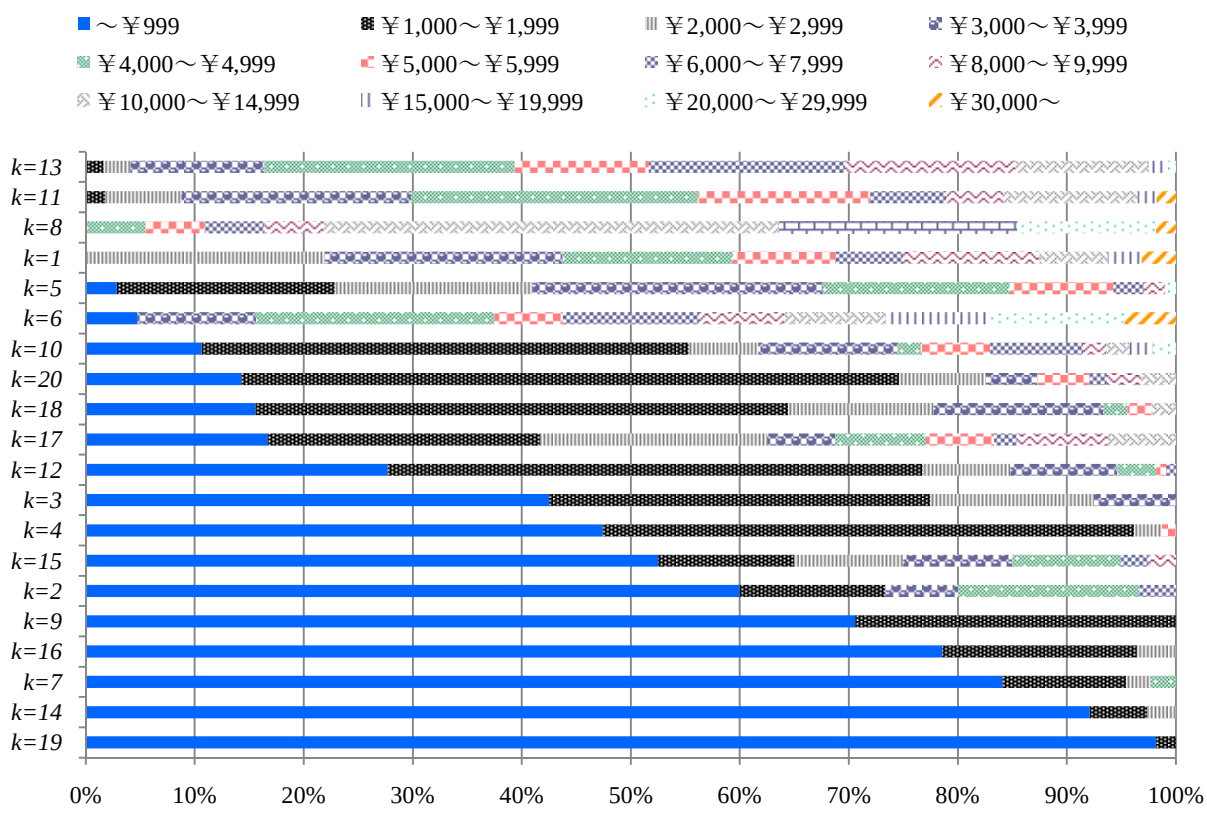


図 4 飲食店分類後の各潜在カテゴリ内での価格帯の分布

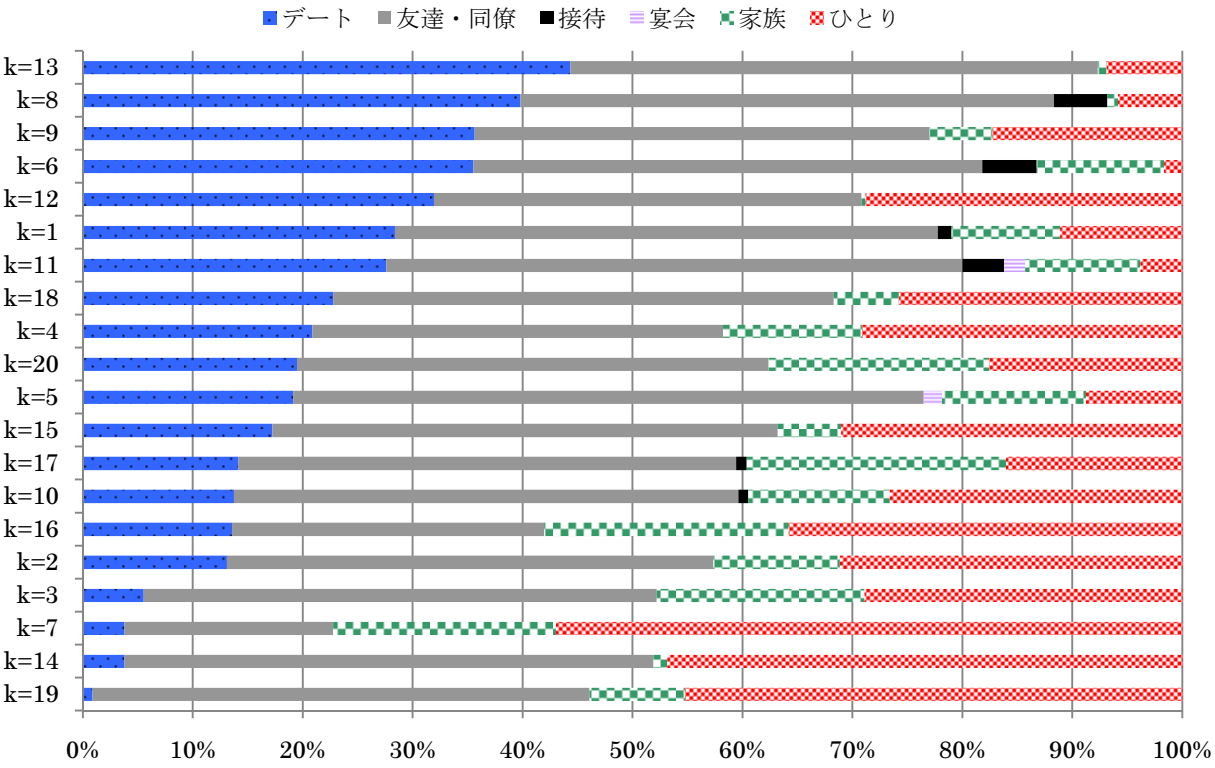


図 5 飲食店分類後の各潜在カテゴリ内での用途の分布