# Q&Aサイトにおける質問文を用いた ソーシャルリスニングの可能性

- Yahoo!知恵袋におけるiPhoneとAndroidの比較から-

吉見 憲二, 谷本 和也, 田中 康裕, 岩井 憲一, 上田 祥二, 針尾 大嗣

## 要旨

本研究では、ソーシャルメディアの投稿データに含まれるノイズやAPI利用環境の変化を問題意識とし、Q&Aサイトの質問文を用いたソーシャルリスニングの可能性を検討した。 具体的には、日本最大のQ&AサイトであるYahoo!知恵袋の公開データを使用し、iPhoneとAndroidカテゴリの質問文の比較を行った。共起ネットワーク分析やトピックモデル、対応分析といった計量テキスト分析の手法を用いた分析結果より、iPhoneカテゴリとAndroidカテゴリのそれぞれにおいてユーザーが抱えている蓋然性の高い「悩み」を把握し、経年的な変化についても一定程度可視化することに成功した。さらに、一連の分析を通して、カテゴリ化された質問文を用いた分析の有用性を示すことができた。

本研究では研究者が取得できる公開データとフリーのソフトウェアの組み合わせという再 現性が高い手法を採用しており、同様のアプローチを別のテーマに幅広く応用することが期 待できる。

キーワード: Q&Aサイト、ソーシャルリスニング、計量テキスト分析、iPhone、Android

#### 1. はじめに

X(旧Twitter)やFacebook、Instagramといったソーシャルメディアの普及に伴い、消費者の生の声をマーケティング等に活用するソーシャルリスニングがさまざまな場面で活用されるようになってきている。しかしながら、こうしたサービスの投稿内容にはノイズが含まれることが多く、有用な知見を得るためには分析上の工夫が必要となることが少なくない。さらに、ソーシャルリスニングのツールとして広く用いられてきたX(旧Twitter)のAPI有料化の範囲が広がったことにより、研究者であっても従来のように気軽に利用することが困難になっている。

こうした背景を踏まえて、本研究ではQ&Aサイトにおける質問文を用いたソーシャルリスニングの可能性に着目する。国立情報学研究所(NII)のデータセット共同利用研究開発センター(DSC)が運営するデータセットの共同利用事業である情報学研究データリポジトリ

(Informatics Research Data Repository: IDR) では、日本最大のQ&AサイトであるYahoo!知恵袋からデータセットが提供されており、研究者であれば申請することで比較的容易に利用することができる。加えて、Yahoo!知恵袋ではQ&Aという形式を採用していることから利用者のより具体的な悩みが質問として顕在化し、質問がカテゴリ化されているため投稿のノイズの影響が比較的軽微であると考えられる。

本研究では、Yahoo!知恵袋におけるスマートフォンカテゴリ内のiPhoneカテゴリとAndroidカテゴリの質問文の分析から利用者の悩みの可視化を試み、その優位点と課題について検証することを目的とする。

## 2. 先行研究

上田ほか (2005) では、テキストのような非構造化データの分析によって携帯電話会社のブランドイメージや自動車会社のブランドイメージを明らかにし、ユーザーからの自由記述による評価が広く活用できることを示している。2000年代後半からのSNSに代表されるソーシャルメディアの普及はこうした方法論の有効性をさらに高め、さまざまな知見が先行研究から得られている。例えば、吉見・樋口 (2012) では、口コミサイトのユーザーレビューを分析し、訳あり商品が支持される要因について検討している。また、ルハタイオパット (2020)では、ソーシャルメディア上のタイ語のデータから、原発事故後の日本の農水産物・食品に対するタイの消費者意識について明らかにしている。

ソーシャルメディアを通して顧客の声を収集,分析し、マーケティング等に活用する試みはソーシャルリスニングとも呼称される。ラパポート(2012)はソーシャルリスニングには、オンライン上のブランドに関するコメントを対象とするソーシャルメディア・モニタリングと自然に発生したオンライン上の会話を対象とするソーシャル・リサーチの2種類のリサーチがあると定義し、前者は戦術的で後者は戦略的なアプローチであると位置づけている。五藤(2012)はソーシャルリスニングによって、「社会のなかに入ってリスニングすることで、消費者や社会をより理解することが可能となり、消費者や社会を中心に商品開発やマーケティング、経営が行われることになる」と指摘している。

このようにソーシャルリスニングの可能性について論じる言説は数多くあるが、ユーザーレビューに代表されるソーシャルデータの分析にあたっては一定の投稿数が求められ、レビューを投稿する行為自体がある程度ロイヤリティの高いユーザーに限定されてしまいやすいという課題がある。長島(2015)はソーシャルメディアによく書き込む人を対象とした調査から、ソーシャルリスニングの課題として、「ネガティブ意見の表明について、躊躇、婉曲化、訂正という傾向が顕著に現れる」ことを挙げている。

さらに、X(旧Twitter)やFacebook、Instagramといったサービスでは、いつでもどこでも

製品やサービスの感想や不満を投稿できる一方で、必ずしも特定の目的のために使用される わけではないため、分析に際して含まれるさまざまなノイズに対応することが求められる。 例えば、渡邉・吉野(2018)はTwitterにおいて1つの投稿で話題が完結しないことから、そ の前後の投稿内容を踏まえて分析する必要性を指摘している。

こうしたソーシャルメディアの課題に対して、Yahoo!知恵袋の投稿データはカテゴリ化された質問文という形式となっているためノイズの影響を受けづらく、実際に困っている利用者が投稿していることから、過度にロイヤリティの高いユーザーに限定されないという特徴を有していると考えられる。佐々木・高橋(2015)では、ソーシャルメディアでも大量に投稿されている「育児の悩み」に関して、Yahoo!知恵袋の質問文を用いることによって、より投稿者の悩みがはっきりした精度の高いデータを取得することに成功している。

吉見ほか(2020)ではこうした観点から、公開されている Yahoo! 知恵袋の投稿データを用いて、iPhone と Androidカテゴリの質問文の比較を行っている。ただし、使用しているデータが3年分で、iPhone に関する投稿が24,957件、Android に関する投稿が7,116件にとどまっている。そのため、今回はデータを追加することで改めて同様の分析を行い、Q&Aサイトにおける質問文を用いたソーシャルリスニングの可能性について検討する。

#### 3. 分析

#### 3.1. 分析に用いるデータ

本研究では、前述した「Yahoo! 知恵袋データ(第3版)」の2019年度から2023年度提供版データより、2014年度から2020年度の7年間の投稿データを使用する。Yahoo! 知恵袋は日本最大のQ&Aサイトであり、多くのユーザーから質問と回答が寄せられている。2023年度提供版データにおける質問数は約217万件、回答数は約559万件となっている。ただし、収録データは収録期間(2018年4月1日 – 2021年3月31日)に投稿され解決した質問の10%がランダムサンプリングされたものである。

分析はスマートフォンカテゴリの下位にあるiPhoneカテゴリとAndroidカテゴリに投稿された質問文を対象とし、最終的に、iPhoneカテゴリで55,470件、Androidカテゴリで14,535件の投稿データを取得した。

## 3.2. 分析方法

分析に当たってはフリーのテキストマイニングソフトウェアであるKH Coder(https://khcoder.net/)を使用した。形態素解析は付属のChaSen(茶筌)を用いているが、一部の辞書に登録されていない用語については強制抽出の対象としている。KH Coderの使用方法や活用事例は樋口(2020)にまとめられている。

具体的な分析手順として、頻出上位語の抽出、共起ネットワーク分析による質問文の可視化、トピックモデルによる話題(トピック)の探索、対応分析による年度間の比較を行った。こうした複合的な分析アプローチから、それぞれの利用者が抱える「悩み」について蓋然性の高い把握を試みる。

なお、各分析では全体の5%以上に登場している単語を対象とした。ここでの単語の登場数は1投稿内に含まれるかを基準としており、1投稿内に複数登場していても重複してカウントはされていない。また、表記ゆれへの対応として英字表記は統一しているが、カタカナ表記については別途抽出した。

#### 3.3. 分析結果

## 3.3.1. 頻出上位語

iPhone と Android の頻出上位25語の名詞をまとめたものが表1である。それぞれ片方にしか登場しない単語については太字で示している。上位10語については概ね共通しているが、iPhoneカテゴリでは、「iTunes」「ID」「SIM」「バックアップ」といった単語が特徴的であった。さらに、「機種」は共通していたものの、「変更」が上位に登場したのはiPhoneカテゴリのみであった。

Androidカテゴリでは、「カード」「SD」「画像」「起動」「ダウンロード」といった単語が特徴的なものであり、カタカナ表記が上位に登場する点はiPhoneカテゴリには見られないものであった。「写真」は両カテゴリに見られるものの、「画像」はAndroidカテゴリにしか登場しないことも興味深い点であった。

#### 3.3.2. 共起ネットワーク分析

続いて、共起ネットワーク分析の結果を図1と図2にそれぞれ示している。共起ネットワーク分析は、対象とする単語が文書内で共通に出現する関係(共起関係)を円と線で可視化したものであり、本分析ではコサイン類似度0.1以上の共起関係を共起ネットワーク描写の基準とした。

iPhoneカテゴリでは、「iTunes」「データ」「バックアップ」といった共起関係や「機種」「変更」と「SIM」や「ID」といった共起関係が頻出上位語の傾向と一致していた。このことは機種変更やデータバックアップが主要な利用者の「悩み」である可能性を示唆している。

一方で、Androidカテゴリでは、「SD」「カード」「データ」や「Google」「インストール」「アプリ」といった単語に共起関係が見られた。これはSDカードへのデータ保存やアプリインストールが主要な利用者の「悩み」である可能性を示唆している。特に、「アプリ」はiPhoneカテゴリと比べても共起している単語数が多かった。

#### 3.3.3. トピックモデル

トピックモデルは「教師なし機械学習」の一種であり、いくつかの単語の集合であるトピックに関して、テキストデータからトピック群を抽出する統計モデルである。トピックモデルの利点として、汎用的な手法であることからその適用範囲の広さが挙げられるが<sup>1</sup>、トピック数を分析者が設定する必要がある。本分析では、トピックモデルの中でも広く用いられている潜在的ディリクレ配分法(LDA)を採用し、トピック数についてはRのIdatuningパッケージより、最適なトピック数を検討した(図3、図4)。最終的に、iPhoneカテゴリは14トピック、Androidカテゴリは8トピックを抽出した。

抽出したトピックの結果は図5と図6の通りとなっている。iPhoneカテゴリでは、「iTunes」によるバックアップ(#4)やSIMの使用(#5)、メールの設定(#6)、機種変更(#14)等に関する質問がトピックとして登場していた。Androidカテゴリでは、アプリのダウンロード/インストール(#2)、SDカードへのデータ保存(#5)、「XPERIA」(#1)と「GALAXY」(#4)に関する質問などがトピックとして登場していた。これらは概ね共起ネットワーク分析の結果と一致するものとなっていた。

## 3.3.4. 対応分析

最後に、対応分析の結果を図3と図4にそれぞれ示している。対応分析は質的データを対象とした主成分分析であり、2次元のグラフ上に単語をマッピングすることで、単語間の関係性を視覚的に表現することができる。さらに、本分析では、各年度のラベルを付記することで、当該のグラフ上に表示している。これにより経年的な変化についても検討が可能である。なお、製品名によって極端な表示になってしまうことから、原点付近を拡大することで視認性を高めている。

iPhoneカテゴリでは横軸の成分1の寄与度が81.49%, Androidカテゴリでは63.57%と高い数値を示しており、年度の変化と一致していることから、経年的な質問傾向の変化が生じていることが読み取れる。加えて、年度のラベルの位置より、質問文の傾向が概ね2年ごとに変化していることがうかがえる。

iPhoneカテゴリでは、「バックアップ」が横軸の中間に位置し、「iTunes」が左側に「SIM」と「ID」がそれぞれ右側に配置されている。このことは、「バックアップ」への関心が経年的な変化に影響されにくい一方で、2016年度以前に悩みが多かった「iTunes」の話題がそれ以降は「SIM」や「ID」に関する話題に変化してきていると解釈することができる。これはiPhoneのサービス展開の流れとも一致している。

<sup>1</sup> トピックモデルを利用した先行研究は、佐藤(2017)、武井ほか(2020)などで挙げられている。

それに対して、Androidカテゴリでは、主要な製品が「XPERIA」から「GALAXY」に変化してきていることが結果から可視化されていた。

## 3.4. 考察

本研究の分析結果より、iPhone ユーザーと Android ユーザーの抱えるそれぞれの「悩み」の一端が可視化されている。一例として、分析対象初期の投稿データでは、iPhone カテゴリにおいて「iTunes」が、Android カテゴリにおいて「SD」「画像」「保存」が特徴的な単語として得られた。これはデータ移行やバックアップにおいてそれぞれのユーザーが抱える困難が質問文に反映された結果だと考えられる。さらに、各端末のサービスの変化によって、こうした「悩み」の経年的な変化が生じていることも確認された。例えば、SIMロックの原則廃止が政策的に決定されたのは2021年10月1日からであるが、それ以前より「SIM」に対する関心が高まっていたことが示されている。こうした一連の結果は、解消された過去のユーザーの「悩み」と新しいユーザーの関心を把握する上で有用な情報となっている。

研究者が取得できる公開データとフリーのソフトウェアの組み合わせから分析を実施し、一定の成果を確認できたことも本研究の意義である。本分析では辞書に登録されていない単語を強制抽出の対象とする処理を行ったものの、それ以外では投稿のノイズに対して特段の対応を行っていない。X(旧Twitter)やFacebook、Instagramといった多くのユーザーを抱えるソーシャルメディアの投稿データをソーシャルリスニングに利用する際に関連性の低いノイズへの対応が必要となるケースが多いことを踏まえると、Yahoo!知恵袋におけるカテゴリ化された質問文はソーシャルリスニングに用いるデータとして精度の高いものであったと結論付けることができる。

ただし、情報学研究データリポジトリ(IDR)におけるデータ公開にはタイムラグがあることから、ソーシャルリスニングの利点であるリアルタイム性を発揮するためには課題がある。現在は事後的な検証しかできないため、マーケティング等の施策に活用するためにはデータを提供するための別の枠組みが必要となるだろう。

さらに、今回は全体の傾向を示すために「全体の5%以上に登場している単語」に分析対象を限定したが、それによって特定の年度に限定された短期的な話題に関して過小評価している懸念がある。新しい製品の登場によって利用者の悩みが変化していくことを踏まえると、トピックの抽出にあたっては期間を限定する方が望ましいかもしれない。対応分析では、2年程度の製品サイクルが示されていたため、分析手法に応じて対象とする期間を変えていくことでより有用な結果が得られる可能性がある。

## 4. まとめと今後の課題

本研究では、日本最大のQ&AサイトであるYahoo!知恵袋を対象に、質問文を用いたソーシャルリスニングの可能性を検討するために、iPhoneとAndroidカテゴリの比較を行った。結果より、iPhoneカテゴリとAndroidカテゴリのそれぞれにおいてユーザーが抱えている蓋然性の高い「悩み」を把握することができた。併せて、先行研究からデータを追加することで、経年的な変化についても一定程度可視化することに成功した。分析のリアルタイム性や投稿量といった課題はあるものの、より明確な問題意識の下に質問が投稿されているQ&Aサイトの質問文を用いてソーシャルリスニングを行う有用性を示すことができた。

本研究は研究者が取得できる公開データとフリーのソフトウェアの組み合わせによって実施されているため、手法の再現性が高いものである。加えて、Yahoo!知恵袋にはスマートフォン以外にもさまざまなカテゴリが存在するため、今後は同様のアプローチをさまざまなテーマに応用することで、引き続きソーシャルリスニングの可能性を検討していきたい。

吉見憲二(成蹊大学経営学部准教授) 谷本和也(佛教大学社会学部講師) 田中康裕(社会データ構造化センター特任研究員) 岩井憲一(滋賀大学教育学部准教授) 上田祥二(成蹊大学経営学部非常勤講師) 針尾大嗣(摂南大学経営学部教授)

#### 謝辞

本研究は ROIS-DS-JOINT 2023 (課題番号:046RP2023, 研究代表者:針尾大嗣) の助成を受けた。また、本研究では国立情報学研究所のIDRデータセット提供サービスによりLINEヤフー株式会社から提供を受けた「Yahoo! 知恵袋データ (第3版)」を利用した。

### 参考文献・URL

- 上田隆穂・黒岩祥太・戸谷圭子・豊田裕貴 (編) (2005) 『テキストマイニングによるマーケ ティング調査』 講談社
- 五藤寿樹 (2012)「社会情報とビジネス:ソーシャルメディアとデータ集約型科学」『社会情報学』1 (2), pp.27-35
- 佐々木裕子・高橋眞理(2015)「インターネットのQ&Aコミュニティサイトにみる $0 \sim 4$ ヵ月児の母親の育児における寝かしつけの悩み テキストマイニングによる分析」『医療看護研究』11(2),pp.28-35

- 佐藤圭 (2017)「マーケティング研究におけるトピックモデルの適用に関する一考察」『経営研究』68(3), pp. 125-148
- 武井美緒·藤野友和·中野純司(2020)「トピックモデルを用いた研究動向の分析」『統計数理』 68(2), pp. 219-231
- 長島直樹 (2015)「ソーシャルリスニングの有効性と限界:書き込みの信頼性と安定性に関する検証実験から」『経営論集』85. pp.11-25
- 樋口耕一(2020)『社会調査のための計量テキスト分析―内容分析の継承と発展を目指して【第 2版】』ナカニシヤ出版
- LINEヤフー株式会社 (2023) Yahoo! 知恵袋データ (第3版). 国立情報学研究所情報学研究 データリポジトリ. (データセット). https://doi.org/10.32130/idr.1.3 (2023年11月20日アクセス)
- 吉見憲二・田中康裕・針尾大嗣・谷本和也・源城かほり・岩井憲一・小舘亮之(2020)「Q&A コミュニティにおける質問文からの製品情報の分析」『情報コミュニケーション学会第 17 回全国大会予稿集』
- 吉見憲二・樋口清秀 (2012) 「共起ネットワーク分析を用いた訳あり市場の考察: 「カニ」と 「ミカン」のユーザーレビューを題材として」 『GITS/GITI research bulletin』 2011-2012, pp.31-39
- ラパポート スティーブン・D(2012)『リッスン・ファースト:ソーシャルリスニングの教科 書』 翔泳社
- ルハタイオパット プウォンケオ (2020) 「海外のSNSデータを用いたソーシャルリスニング ―輸出に向けた対象国の消費者意識・ニーズの把握―」『関東東海北陸農業経営研究』 110, pp.5-13
- 渡邉小百合・吉野孝(2018)「観光地名なしツイートからの観光地に関する感想の抽出手法」『情報処理学会論文誌』59(1), pp.43-51

表1 頻出上位語

順位		iPhone		Android				
川貝汀丛	抽出語	登場数	割合	抽出語	登場数	割合		
1	iPhone	28948	52.2%	Android	5252	36.1%		
2	アプリ	10858	19.6%	スマホ	4911	33.8%		
3	お願い	9594	17.3%	アプリ	4461	30.7%		
4	方法	7107	12.8%	お願い	2875	19.8%		
5	画面	5830	10.5%	XPERIA	2467	17.0%		
6	質問	5786	10.4%	方法	2344	16.1%		
7	使用	5586	10.1%	画面	2015	13.9%		
8	Apple	5425	9.8%	機種	1968	13.5%		
9	機種	5236	9.4%	使用	1630	11.2%		
10	購入	4969	9.0%	設定	1547	10.6%		
11	設定	4650	8.4%	質問	1522	10.5%		
12	場合	4509	8.1%	カード	1507	10.4%		
13	iTunes	4381	7.9%	Google	1487	10.2%		
14	写真	4305	7.8%	表示	1486	10.2%		
15	iPhone6	4294	7.7%	SD	1276	8.8%		
16	変更	4048	7.3%	画像	1217	8.4%		
17	ID	4026	7.3%	データ	1178	8.1%		
18	表示	3921	7.1%	GALAXY	1074	7.4%		
19	データ	3915	7.1%	端末	1071	7.4%		
20	SIM	3543	6.4%	購入	1050	7.2%		
21	メール	3479	6.3%	アンドロイド	1006	6.9%		
22	バックアップ	3365	6.1%	ダウンロード	913	6.3%		
23	iOS	3307	6.0%	起動	904	6.2%		
24	スマホ	3128	5.6%	au	874	6.0%		
25	自分	3069	5.5%	写真	861	5.9%		

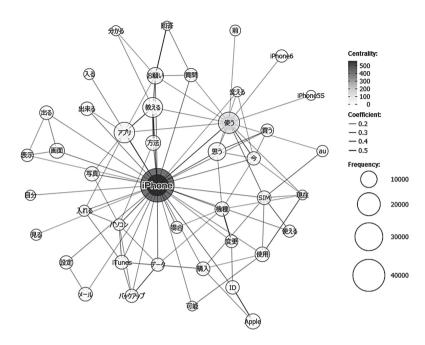


図1 共起ネットワーク分析結果 (iPhone)

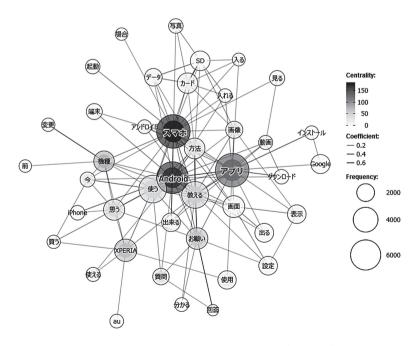
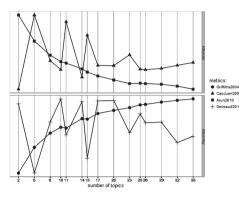


図2 共起ネットワーク分析結果 (Android)



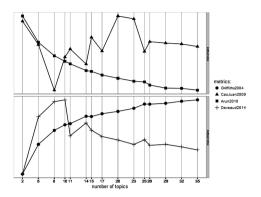


図3 トピック数の指標 (iPhone)

図4 トピック数の指標 (Android)

#1	了	#2		#3	3		#4			#5		#6		#7	
使う	0.468	アプリ	0.664	教	₹ें	0.236	写真		0.258	使用	0.29	5 メール	0.287	思う	0.340
今	0.206	ios	0.184	方法	去	0.218	iTun	es	0.238	SIM	0.26	設定	0.267	買う	0.234
携帯	0.162	教える	0.133	iPh	one5	0.147	バック	アップ	0.218	au	0.17	ち 教える	0.225	使う	0.213
変える	0.139	方法	0.016	<del>/)</del> /	かる	0.146	パソコ	ン	0.162	使える	0.14	5 方法	0.103	iPhone5S	0.139
前	0.023	出る	0.000	状	II.	0.137	データ	,	0.123	現在	0.08	3 お願い	0.090	使える	0.061
#9		#10		#11	$\int_{-}^{-}$			#12	2 _/	_		#13		#14	
画面	0.326	購入	0.277	iPhone	0.9	39		iPho	ne 0.	889		出来る	0.272	機種	0.303
出る	0.269	お願い	0.195	可能	0.0	52		デー	9 0.	108		iPhone6	0.224	変更	0.240
表示	0.222	質問	0.188	教える	0.0	03		状態	0.	001		入れる	0.181	iPhone7	0.154
見る	0.177	スマホ	0.170	お願い	0.0	02		教え	る  0.	001		入る	0.175	思う	0.094
敗える	0.003	回答	0.121	携帯	0.0	02		写真	;  o.	000		前	0.116	場合	0.091

図5 トピックの抽出結果 (iPhone)

#1		#2			#3		#4	
XPERIA	0.257	アプリ	0.590	)	画面	0.238	お願い	0.132
機種	0.215	出来る	0.17	5	出る	0.174	端末	0.124
使用	0.167	ダウンロー	- F 0.10	1	表示	0.171	質問	0.109
思う	0.124	インストー	ル 0.092	2	設定	0.171	購入	0.108
au	0.092	入れる	0.018	3	見る	0.102	GALA	0.107
#5		#6		#7			#8	
カード	0.199	教える	0.220	Android	0.537		スマホ	0.586
SD	0.184	方法	0.213	使う	0.198		使う	0.197
データ	0.144	Google	0.180	教える	0.145		前	0.072
画像	0.137	お願い	0.097	iPhone	0.106		今	0.068

図6 トピックの抽出結果 (Android)

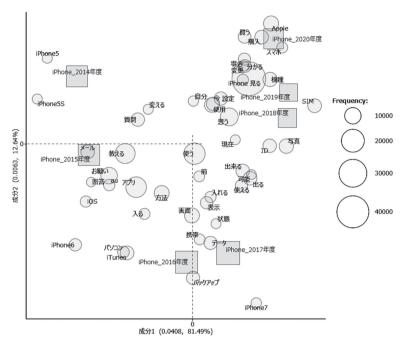


図7 対応分析結果 (iPhone)

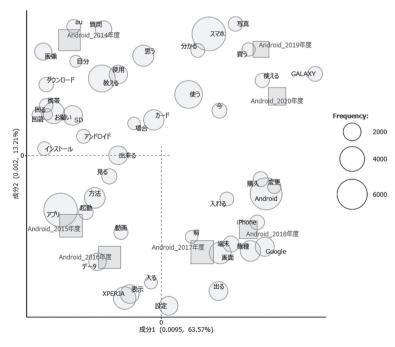


図8 対応分析結果(Android)