Androidアプリにおけるカテゴリ分類法の検証
---K-means法を用いたアプローチ---

お茶の水女子大学
理学部　情報科学科　金子研究室

永森　枝里子

目次

1.背景

- 膨大なデータを管理するために、カテゴリ分類は必須
- 人の手で正確に行おうとすると膨大な時間がかかる
- 人の手で行っても、分類者の主観が入ってしまい、基準を統一するのは非常に難しい
- Apple App Storeとは違いAndroid Marketは審査がないため、1日に膨大な量のアプリが追加される

自動で簡単にカテゴリ分類する手法が必要

K-means法を用いたクラスタリングによるカテゴリ分類法を検証

2.実験手順（1）

1. Androidアプリのデータ（名前、説明文）をWebサイトから収集する

2. 各アプリについて任意のIDを振り分け、実験結果との比較用に手作業でカテゴリを分類し、以下の形のデータを作成する
   ［ID、名前、カテゴリ、説明文］

3. Text Mining Studioを用いて説明文を形態素解析にかける

4. 形態素解析の結果から単語の有無一覧（マトリクス）を生成し、Visual Mining Studioを用いてK-means法によりクラスタ分析を行う

5. クラスタリングした結果と、手で分類したカテゴリを比較し、クラスリングによるカテゴリ分類の精度を検証する

2.実験手順（2）

- テキストマイニングの問題点
  - ゴミが出やすい
  - 必ず正しい位置で単語を分割できるとは限らない

- まずはできるだけ手を加えない状態で実験
- その後上記問題点による影響を考え、精度を上げることを検討する

3.使用データ

- データ収集Webサイト

  Android Market（http://www.**android**.com/market）

  ドコモマーケット（http://www.dcm-gate.com/）

- 対象データ

  日本語の説明のあるアプリ1271件

- 説明文について

  今回は先頭から100文字目までを抽出

4.実験準備：カテゴリ分類

- 大カテゴリ（11項目）と小カテゴリ（42項目）を用意
- 大カテゴリの分け方に近い結果を目標とする

大カテゴリ	小カテゴリ
エンターテイメント	エンターテインメント
ゲーム	RPG・シミュレーション・その他
	カジノ・カード
	シューティング・スポーツ
	パズル・ボードゲーム
	アクション・アドベンチャー
ツール	音量マナー・各種開発／計測等
	他デバイス連携
	スクリーン・ライト設定
	セキュリティ
	その他ユーティリティ
	ファイル・メモリ・電池管理
	起動・動作補助・トグル
	時計・タイマー
	電話帳管理
	文字入力
	インストール・アンインストール関連
ネット関連	RSS・ブックマーク
	コミュニティ/SNS
	ショッピング
	ブラウザ・ネットワーク
	各種検索ツール

大カテゴリ	小カテゴリ
音楽	音楽
画像・カメラ・動画	アイドル・グラビア画像
	カメラ・撮影・画像編集
	動画
仕事・勉強	スケジュール・カレンダー
	タスク管理
	メモ・レコーダー
	辞書・学習・電卓・計算機
情報	ニュース
	医療・健康・美容
	金融
	地図・ナビ
	鉄道・道路情報
	天気
	旅行・グルメ
通話・メール・接続	wi-fi・接続環境／通話・チャット等
	メール
電子書籍	書籍・コミック・新聞
壁紙・着信	アイコン・スキン・テーマ
	着信・壁紙管理等

5.【実験1】
テキストマイニングの結果をそのまま利用(1)

- Text Mining Studioで形態素解析
  - データの取り込み～分かち書きの実行

    ＜投入データ＞
    - ［ID，名前，カテゴリ，説明文］

    ＜設定＞
    - 説明文を「テキスト」に設定
    - **分かち書きのみ**（形態素解析のみを行いたいため）
    - **字面を合わせ込む**（大文字小文字、全角半角の表記ゆれを改善）

  - 単語頻度解析

    ＜設定＞
    - フィルタ条件を**名詞（一般名詞）のみ**にする（カテゴリを分類する際に、多くの場合一般名詞を見てカテゴリを判断しており、一般名詞のみで十分だと判断したため）
    - 出現回数**2回以上**のものを全て出力
    - 単語を全選択し、「単語有無表の作成」でマトリックスを作成

5.【実験1】
テキストマイニングの結果をそのまま利用(2)

- Visual Mining Studioでクラスタリング
  - データの取り込み
    ＜投入データ＞
    - Text Mining Studioで作成したマトリックスのデータ
  - K-means法によるクラスタリング
    ＜設定＞
    - 対象列には、TMSで出力された単語を**全て**選択
    - 距離計算：**Cosine**
    - クラスタ数：**11**（大カテゴリ数）
    - 繰り返し回数：100
    - 初期値：**手動「1」**

5.【実験1】
テキストマイニングの結果をそのまま利用(3)

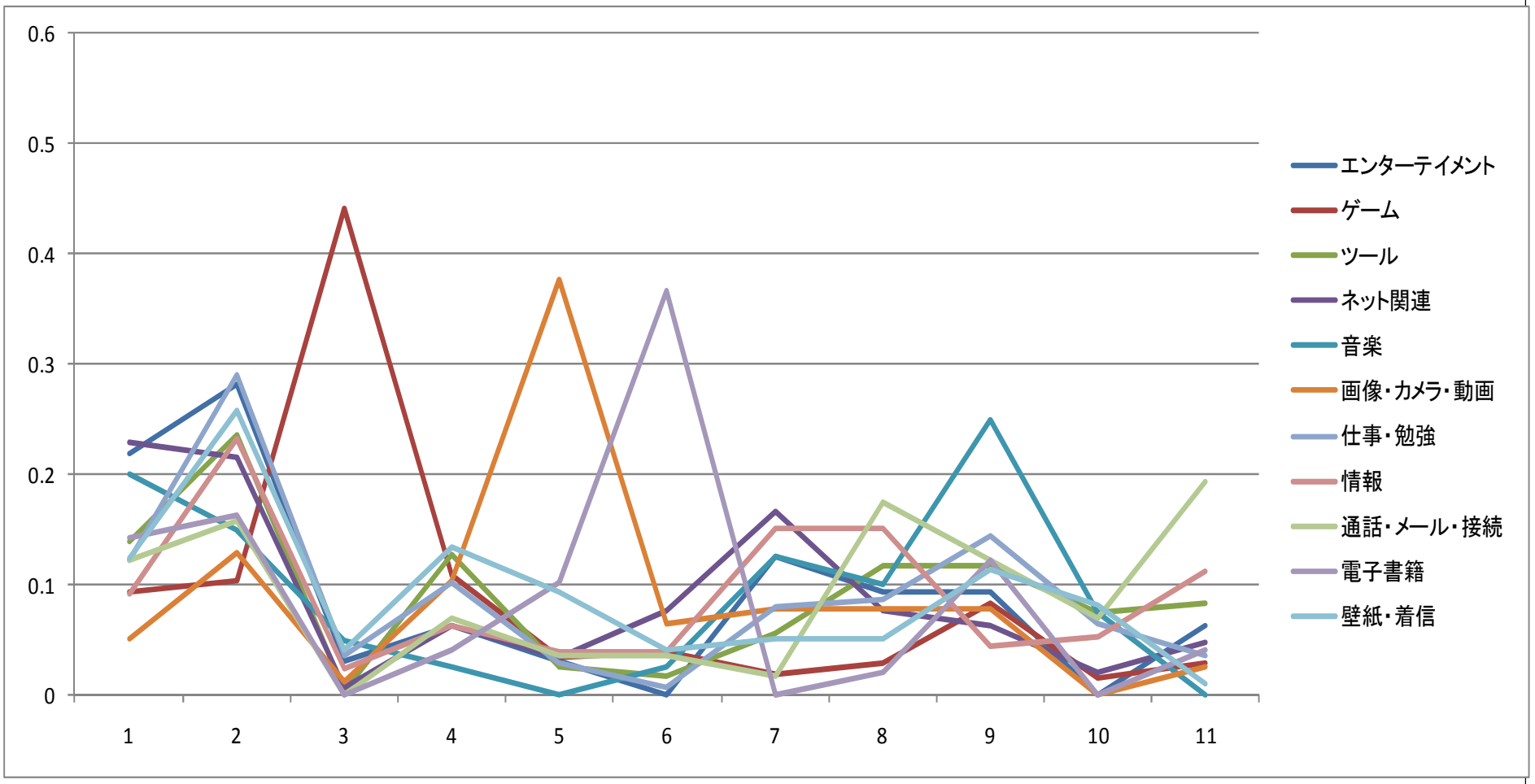
- クラスタリングの結果
  - Cluster infoの値が大きい単語を、そのクラスタの特徴を表す単語として抽出（図1）
  - 最初に分割した大カテゴリとクラスタの比較
    カテゴリごとの各クラスタが占める割合を計算し、グラフ化（図2）

［カテゴリXに対するクラスタAの割合］
＝［カテゴリX内のクラスタAの数］／［カテゴリX全体の数］

図1　特徴を表す単語抽出（実験1）

クラスタID	頻度Rank 1	2	3	4	5
1	Android	アプリ	用	端末	アプリケーション
2	アプリ	い	画面	情報	Xperia
3	ゲーム	的	人	世界	数字
4	い	無料	版	画像	日本語
5	写真	アプリ	画面	バージョン	カメラ
6	中	上	他	文庫	青空
7	情報	最新	Google	ブラウザ	http
8	アプリケーション	情報	履歴	い	他
9	v	画面	音	へ	ボタン
10	ウィジェット	時計	画面	Ver	ホーム
11	時	表	時刻	複数	ツール

図2　カテゴリごとの各クラスタが占める割合（実験1）



5.【実験1】
テキストマイニングの結果をそのまま利用(4)

- 考察
  - キーとなる単語（図1赤字）のわかりやすいカテゴリ「ゲーム、画像・カメラ・動画、電子書籍」は、テキストマイニングの結果をそのまま利用してもクラスタリングでうまく分かれやすい
  - 図1の青字のように「Android、アプリ、アプリケーション」といったどのアプリに入っていても良さそうな特徴を表さない単語がキーとなって分割されてしまっている可能性がある
    ⇒これらの単語を分析の対象から除く
  - 「アプリ・アプリケーション」のような字面を合わせ込むだけでは類義語として判定されない表記ゆれがある
    ⇒自分で辞書を作成

6.【実験2】
類義語辞書の利用と分析対象の検討（1）

- 類義語辞書登録
  - 辞書の作成
    - 単語頻度解析の一覧から類義語に当たるものを一覧としてまとめる（csv形式）

      ＜例＞

App	名詞　一般	アプリ	アプリケーション	アプリケーショ	アプリケ
amazon	名詞　一般	アマゾン			
Android	名詞　一般	アンドロイド			
internet	名詞　一般	インターネット	net	ネット	
インターフェ	名詞　一般	インターフェース			
Wiki	名詞　一般	ウィキ	Wikipedia		

  - 辞書の読み込み
    - Text Mining Studioで作成した辞書を読み込み、「置換処理」で前回の解析結果に反映させる
    - 実験1と同様に単語有無一覧のマトリックスを作成する

6.【実験2】
類義語辞書の利用と分析対象の検討（2）

- Visual Mining Studioでクラスタリング

  ＜設定＞

  - 対象列には、分析の参考にならない以下の単語を除く
    ［App，Android，Ver，v］
  - その他の設定は前回と同様

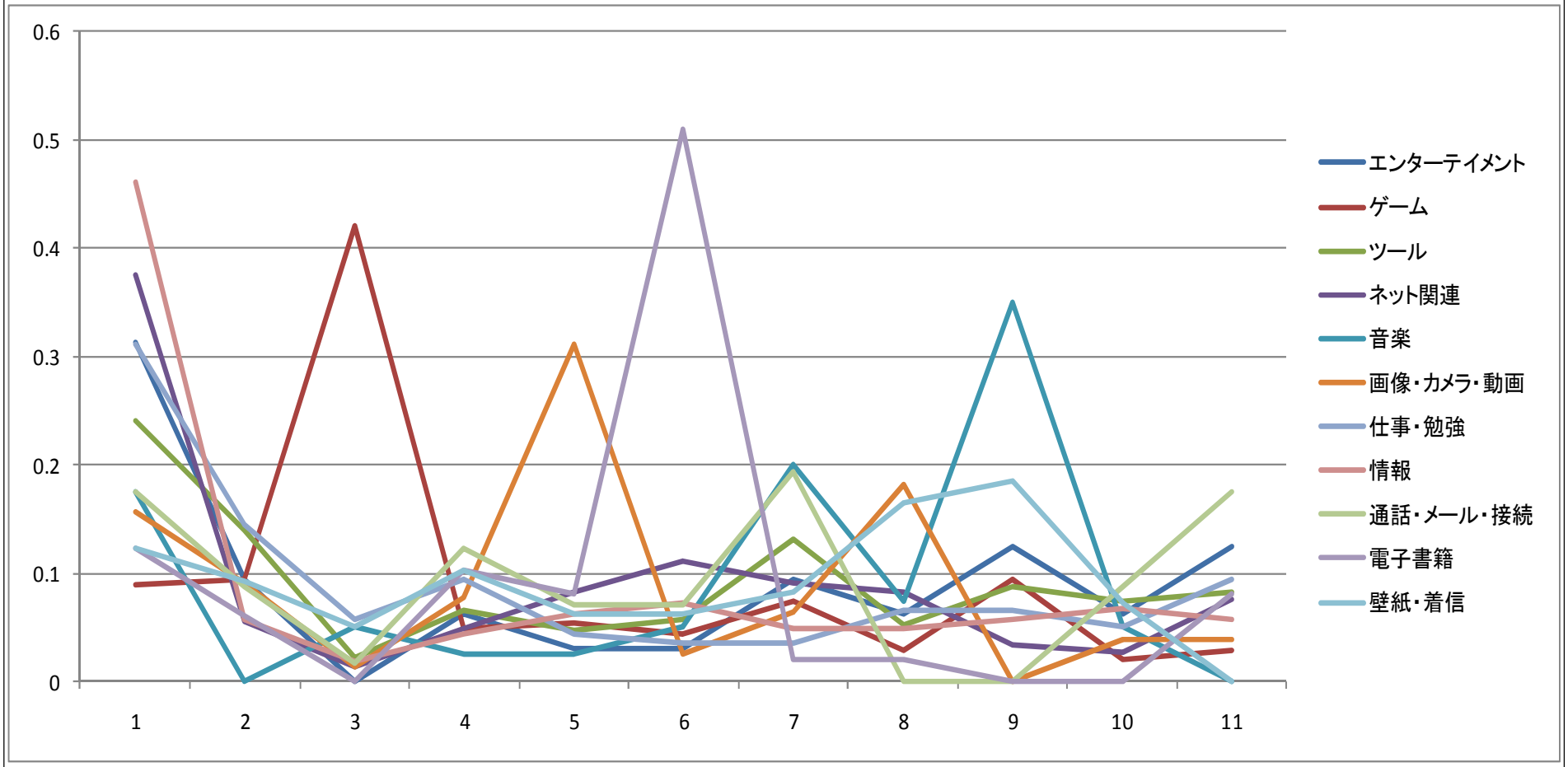
- クラスタリングの結果

  - Cluster infoの値が大きい単語を、そのクラスタの特徴を表す単語として抽出（図3）
  - 最初に分割した大カテゴリとクラスタの比較
    カテゴリごとの各クラスタが占める割合を計算し、グラフ化（図4）
  - 実験1の結果と比較する

図3　特徴を表す単語抽出（実験2）

クラスタID	頻度Rank 1	2	3	4	5
1	情報	い	Web	用	Simeji
2	画面	HOME	Widget	Color	ボタン
3	ゲーム	人	的	数字	世界
4	Free	版	Data	有料	的
5	写真	履歴	い	中	自動
6	中	上	他	文庫	internet
7	Phone	Mobile	情報	音	Email
8	image	カード	SD	壁紙	ファイル
9	音	ボタン	種類	友達	着メロ
10	Widget	情報	時計	Color	特徴
11	時	表	Email	情報	時刻

図4　カテゴリごとの各クラスタが占める割合（実験2）



6.【実験2】
類義語辞書の利用と分析対象の検討(3)

- 考察
  - テキストマイニングのデータをそのまま利用した実験1と比較すると、うまく分かれたカテゴリの種類に大きな変化はないが、全体的に割合が大きくなっており、カテゴリとクラスタの関係性が強まったことが分かる
  - クラスタ1は情報カテゴリの割合が大きいが、その他のカテゴリの割合も大きく、キーとなる単語がわかりにくくどこに分類してよいかわからないものが集まってしまったように思う
  - クラスタ4は図4のグラフでは特に特徴がないように見えるが、図3を見るとわかるように、無料のアプリ(またはそのアプリの有料版)が多く集まっている
- さらに精度を上げるために…**相関の強い単語をグループ化**

7.【実験3】
相関の強い単語をグループ化（1）

- グループ化することの利点
  - 相関の強い単語をグループ化して一つとみなすことで、他の単語との相関関係がより重視されるようになる

- 相関係数を求める
  - データの作成
    - 単語の有無一覧を**リスト形式**にする ［ID，単語］
  - Visual Mining Studioで単語間の相関係数を求める

    ＜投入データ＞
    - 上記リスト形式のデータ

    ＜設定＞
    - 変数1、変数2に単語を**全て**選択

7.【実験3】
相関の強い単語をグループ化（2）

- 単語のグループ化
  - 相関係数0.6以上の単語をピックアップし、グループ化する

    ＊単語A,B,Cについて

    相関係数（A,B）＝0.8　相関係数（B,C）＝0.7　相関係数（A,C）＝0.5の場合、AとCの相関係数は0.6未満だが、それぞれBとの相関が0.6以上なので全て同一グループとみなす

  - グループの中から代表語を一つ決め、辞書として扱う

＜例：赤字が代表語＞

G1	（株）	EMCOM	**FX**	チャート	レート	為替	経済	口座	指標	証券
G2	6つ	R	TOEIC	マスタリー	**英単語**	熟語	本体			
G3	internet	オススメ	国内	作品	図書館	青空	電子	文学	**文庫**	
G4	**mixi**	コミュニティ	トピ	フォト	マイミク	日記				

7.【実験3】
相関の強い単語をグループ化（3）

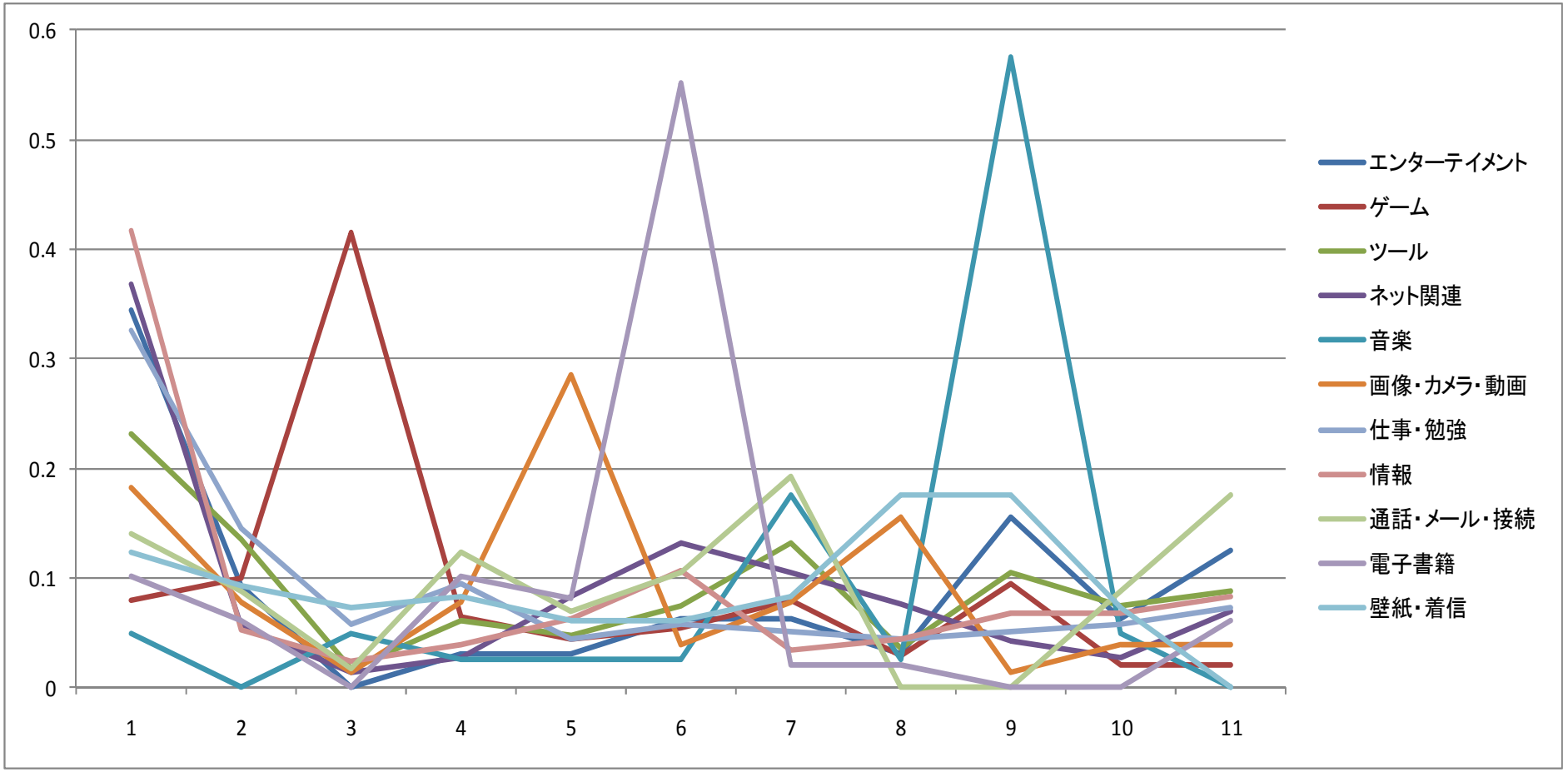
- 辞書登録～形態素解析～クラスタリング
  - 実験2と同様に行う

- クラスタリングの結果
  - Cluster infoの値が大きい単語を、そのクラスタの特徴を表す単語として抽出（図5）
  - 最初に分割した大カテゴリとクラスタの比較
    カテゴリごとの各クラスタが占める割合を計算し、グラフ化（図6）
  - 実験1、実験2の結果と比較する

図5　特徴を表す単語抽出（実験3）

クラスタID	頻度Rank 1	2	3	4	5
1	情報	い	用	Web	Simeji
2	画面	HOME	Widget	Color	ボタン
3	ゲーム	人	世界	的	数字
4	Free	版	有料	Data	的
5	写真	履歴	い	自動	クリック
6	文庫	中	上	他	japan
7	Phone	Mobile	情報	的	音
8	image	壁紙	サイズ	ファイル	情報
9	音	特徴	ボタン	種類	Music
10	Widget	情報	時計	Color	特徴
11	時	表	時刻	Email	い

図6　カテゴリごとの各クラスタが占める割合（実験3）



7.【実験3】
相関の強い単語をグループ化（4）

- 考察
  - 図6のグラフの形は実験2とほぼ同じ
  - 実験2と比べて、音楽カテゴリのクラスタ9の割合と電子書籍カテゴリのクラスタ6の割合が大きく伸びている

    単語を見てもMusic、文庫といったキーとなる単語が実験2よりも上位にきていることがわかる
  - クラスタ1への偏りはあまり改善されなかった

8.まとめと今後の展望

- まとめ
  - 形態素解析とK-means法を用いて分類したカテゴリを、人の手で分類したカテゴリに近づけるように検証した
  - キーとなる単語が決まっている場合はある程度の精度を持って分類できる
  - 形態素解析を行う際に、辞書機能を用いることで、分類の精度を上げることができる

- 今後の展望
  - 今回はクラスタ数を大カテゴリ数の11のみで行ったが、実際にクラスタリングでカテゴリ分類を行うのであれば、クラスタ数も検討する必要がある
  - 単語の類似度を測ることで、キーとなる単語がない場合にもうまく分かれるようにしたい

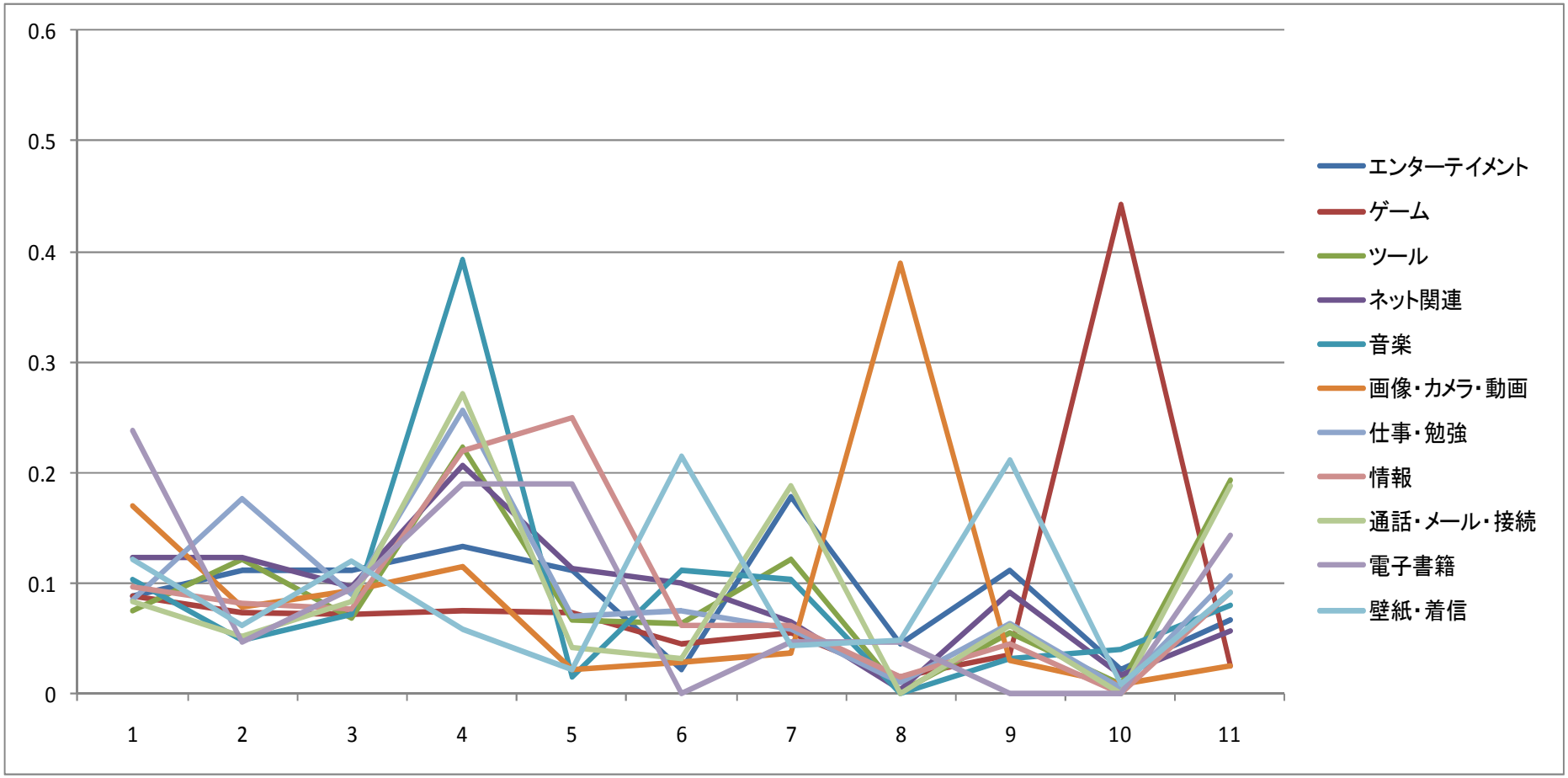
9.番外編：英文を用いた検証（1）

- 使用データ
  - 対象データ
    英文の説明のあるアプリ2604件
  - 説明文について
    今回は先頭から100文字目までを抽出
- 形態素解析～クラスタリング
  - 英文だと単語頻度解析で品詞による指定ができないようなので、全単語を出力
  - クラスタリングの際に以下のような目立った代名詞、前置詞、動詞を対象列からはずした
    例）the, a, is, you, your, for, of, this, in, with, on, it, from…
  - その他は日本語と同様

図7　特徴を表す単語抽出（英文）

クラスタID	頻度Rank 1	2	3	4	5
1	New	version	Fixed	browser	photo
2	can	simple	use	not	easy
3	will	update	all	fix	D
4	search	free	allows	music	Features
5	up	news	world	Google	client
6	have	theme	must	If	don
7	phone	into	battery	Turn	screen
8	sexy	girls	wallpaper	girl	so
9	Open	home	full	use	need
10	game	puzzle	classic	play	fun
11	widget	screen	home	add	clock

図8　カテゴリごとの各クラスタが占める割合（英文）



9.番外編：英文を用いた検証（2）

- 考察
  - 日本語と同様、キーとなる単語がわかりやすいようなカテゴリ「ゲーム、画像・カメラ・動画、音楽」はうまく分かれやすい
  - アルファベットしかない分、日本語よりも表記ゆれが少ないと思われるので、辞書を使わなくてもある程度精度は上げられるのではないかと思う
  - Text Mining Studioに英和辞書を入れて、日本語と英語が混ざっていてもカテゴリ分けができるか検証してみたい