“商品カテゴリ”および“取扱店舗”の統計情報を用いた商品タイトルに含まれるフレーズの重要度判定

05T4074A
久保田　敦

目次

1. はじめに
2. 研究背景
3. 提案手法
4. 精度評価
5. おわりに

1.はじめに

- 研究の目的

  商品名寄せシステムの精度向上。

  名寄せ・・・表記の異なる同一データを
  同じデータと見極める技術

- 研究内容

  商品タイトル中のフレーズ（部分文字列）の重要度を判定する手法の開発。

  今回はそのフレーズ認識手法の説明。

2.研究背景

- オンラインショッピングサービス問題点

  複数のショップから商品を検索する時に名寄せが必要

  {26才，26歳，二十六歳}
  {斉藤，斎藤，サイトウ}　←ショップによって書き方が異なる
  {光学マウス，光学式マウス}

- 解決策

  商品名の文字列をフレーズに分割しその類似度で同一商品かを判断。

宣伝などの商品に関係ない重要度の低い情報（フレーズ）を事前にテキストから取り除く必要がある。

3.提案手法

3.1 用語の定義

用語		説明	例
商品タイトル		商品名の文字列	今なら送料無料！パルック蛍光灯クール色
ストアID		店舗毎の整数	1, 2, 3, ・・・
カテゴリID		カテゴリ毎の整数	1, 2, 3, ・・・
フレーズ		商品タイトル内の部分文字列	送料無料,パルック蛍光灯
ラベル	ノイズ	判別に関係ないフレーズ	送料無料
	シード	判別に有効なフレーズ	パルック蛍光灯,クール色
	中立フレーズ	ノイズでもシードでもない	

3.2　機械学習によるラベル判定

システムの目的は商品タイトルに含まれる全フレーズに対しラベル判定を行い、商品タイトル含まれるノイズとシードを認識すること。

その判定に機械学習の二値分類を用いる。

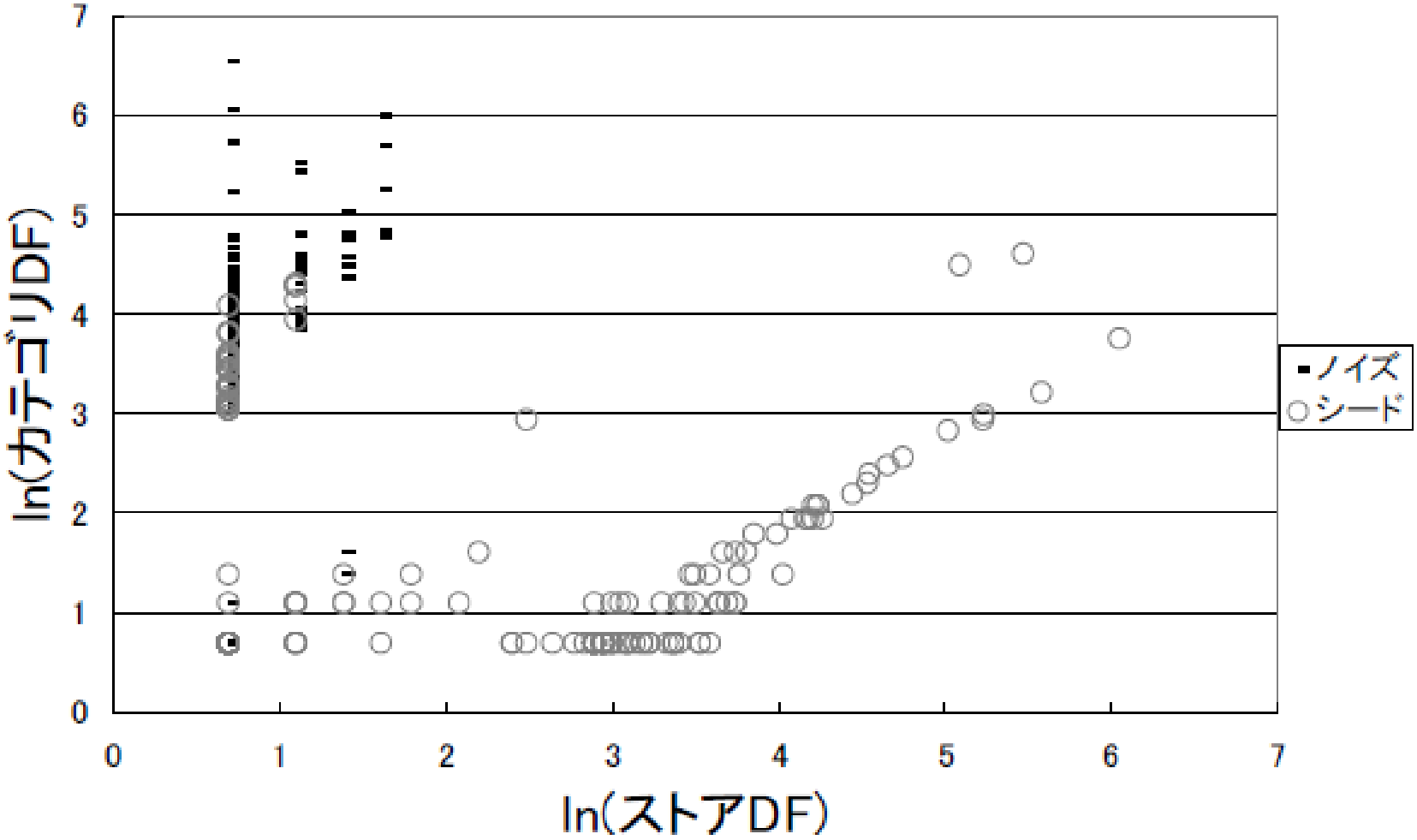
3.3　ストアDF、カテゴリDF

提案：フレーズが出現する商品データのストアIDとカテゴリIDの頻度情報の利用

用語	説明	予想	
		ノイズ	シード
ストアDF	フレーズが使われているストアIDの数	低	高
カテゴリDF	フレーズが使われているカテゴリIDの数	高	低

■ ストアDF、カテゴリDFに関するフレーズの散布図

Yahoo!ショッピング中の商品データに含まれるフレーズから、人手でノイズとシードを収集した。



■ フレーズのストアDF、カテゴリDFの例

フレーズ	ラベル	ストアDF	カテゴリDF
/3500円/以上/の/お買い上げ/	ノイズ	1	131
/送料/無料/セール/	ノイズ	4	148
/岩盤浴/マット/	シード	146	34
/超音波/加湿器/	シード	95	8

[ストアDF、カテゴリDF] と [ノイズ、シード] の関係が見出せたよって、[ストアDF,カテゴリDF] を機械学習の素性として用いることは有効であると考えられる。

4.精度評価

4.1 ラベル判定のガイドライン

フレーズをカテゴリに分け各カテゴリに属するフレーズにラベル判定するためのガイドライン。中立フレーズは扱わない。

カテゴリ	ラベル	フレーズ例
ブランド	シード	/アル/ベロベロ/
形状・容量	シード	/ミニ/ボトル/5ml/
社名（製造）	シード	/日立/
社名（製造・小売）	シード	/カワセ/
社名（小売）	ノイズ	/浪花/商人/
商品カテゴリ	ノイズ	/モダン/アジアン/家具/
人名	ノイズ	/掛布/雅之/
ラベル混合	中立	/送料/無料/ドルチェ/＆/ガッバーナ/
未完結フレーズ	中立	/税抜/3000円/以上/で/
イベント	中立	/東急/フード/ショー/

4.2 学習データの違いによる評価

■ 学習データを2種類用意しての精度比較

データA：Yahoo頻出（人手判定）＋ wikipedia自動収集フレーズデータB：分離平面から遠い500件ずつをラベル判定したフレーズ

- A, Bを用いてランダムに収集した300件のフレーズのラベル判定の結果を評価する

各種データ名	件数（ノイズ数）	
学習データA	2428	(598)
学習データB	603	(134)
テストデータ	286	(99)
Yahoo! 商品データ	5,848,419	
Yahoo! フレーズ	4,831,038	

素性は**フレーズDF比**（カテゴリDFに対するストアDFの比率）

■ 評価結果

学習データ	ノイズ判定精度	シード判定精度
学習データA	0.61	0.93
学習データB	0.62	0.92

- 学習データによらずシード判定精度が良い。
- ノイズの判定精度はシードに比べ低い。

- 本システムから質の良いシードリストを自動生成できる。
- ノイズ判定されたフレーズのうちスコアの高いものから高品質なノイズリストは作成できる。

4.3 素性の違いによる評価

フレーズfを含む商品タイトルの総数をfのアイテムDF
fを含む全商品タイトルの文字列長の平均値をfの平均タイトル長

- これらとフレーズDF比を組み合わせた場合の精度比較を行う。

素性ID	名称
1	フレーズDF比
2	アイテムDF
3	平均タイトル長

素性ID	ノイズ判定精度	シード判定精度
1	0.62	0.92
2	0.00	0.66
3	0.00	0.66
1,2	0.60	0.92
2,3	0.00	0.66
1,3	0.00	0.66
1,2,3	0.40	1.00

結果からフレーズDF比を単独で用いたほうが最も制度が良い。

5.おわりに

- まとめ
  - 高品質なシードリストの自動生成
  - ノイズ判定結果を用いたノイズ作成の効率化

- 今後の課題
  - ノイズ判定精度の向上
  - 中立フレーズの扱いについて