“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

前澤 敏之　山下 達雄　荻原 由紀恵
ヤフー株式会社
{tmaezawa,tayamash,yogihara}@yahoo-corp.jp

1　はじめに

名寄せとは表記の異なる同一データを同定する自然言語処理技術の応用である。とくに商品名寄せは質のよい商品検索を実現するには不可欠である。我々はテキスト間の類似度をベースとした商品名寄せシステムの精度向上を目的として、商品タイトル中のフレーズの重要度を判定する手法を開発した。本稿では我々が開発したフレーズ認識手法を提案する。

2 節では研究背景としてテキスト間の類似度を用いた商品名寄せに関する問題点を述べ、3 節では問題を解決するための手法を提案する。4 節では提案手法に基づき開発したシステムの評価結果を報告する。

2　研究背景

オンラインショッピングサービスではユーザの利便性向上のために、同一商品をあわせて提示する機能 (商品名寄せ) や類似商品検索の機能が求められる。類似商品検索における判定条件を最も厳しくして同一商品同士のクラスタリングに限定したものが商品名寄せであると考えると、これらは同じ枠組みで扱うことができる。

我々は名寄せを「住所録などのデータベースの各レコードが持つフィールドをキーとしてマッチングを行い、キー一致率の高いレコードを同一レコードとする処理」として扱う。このとき同じ内容のキーであっても表記が異なる場合がある。例えば

{26 才, 26 歳, 二十六歳 }
{ 斉藤, 斎藤, サイトウ }
{ 光学マウス, 光学式マウス }

などはそれぞれ同じものを指すキーの集合である。こうしたキーは文字列の完全一致によるマッチングを行うことができない。このため名寄せにおいてはキー同士の類似度を求めることが重要な問題となる。

例えば Lee ら [3] はキー文字列をトークンに分割しトークンの一致率で類似度を計算している。また Aizawa ら [1] は Suffix array-based blocking という Suffix array[4] を用いたトークンベースの手法を提案している。

我々は商品データを名寄せするに当たってキーとして商品タイトルを利用する。このとき商品タイトルには商品に対する宣伝表現が含まれている場合がある。例えば

> 1 月中旬発送予定！！美味しい♪バストアップクッキー「F カップクッキー」（代引手数料無料！ 3 個以上で送料無料！）

という商品タイトルにおいて「1 月中旬発送予定！！」や「代引手数料無料！」などの文字列は商品を宣伝するための表現であり商品自体の認識には寄与しない。類似度を適切に計算するためには、このような重要度の低い情報を事前にテキストから取り除く必要がある。ウェブから収集した商品データにもこうした情報は含まれており名寄せにおける大きな問題となっている。

一方で JAN コードなどの商品識別コードや工業製品の型番など、商品識別の際に非常に有効な文字列もある。こうした重要度の高い情報を優先的に利用することで適切な類似度計算を行うことができる。

3　提案手法

2 節で述べたように適切な名寄せを行うには商品タイトルから商品識別に寄与する部分のみを抽出することが重要である。我々は商品識別における重要度に応じて商品タイトルの部分文字列をノイズあるいはシー

ドとして識別する手法を提案する。本節では提案手法について説明を行う。

3.1 用語の定義

我々が対象とする商品データは表 1 で示される構造を持つ。**商品タイトル**とは各商品データに与えられたタイトル文字列である。**ストア ID** とは商品データを扱う店舗ごとに一意に与えられた整数である。**カテゴリ ID** とは商品データが属するカテゴリごとに一意に与えられた整数である。

フレーズとは商品タイトルに含まれる部分文字列であり「商品タイトル中の名詞形態素連続」もしくは「記号を含まない形態素連続」として定義する。フレーズはひとつの商品タイトル中に複数含まれ重複も認める。例えば商品タイトル

今なら送料無料！パルック蛍光灯　クール色

におけるフレーズは表 2 および表 3 のとおりである。“/” で区切られた部分がひとつの形態素をあらわしている。

表 1　商品データの構造

フィールド	データ型
商品タイトル	string
ストア ID	integer (≧ 1)
カテゴリ ID	integer (≧ 1)

表 2　商品タイトル中の名詞形態素連続

/今/
/送料/無料/
/パルック/蛍光灯/
/クール/色/

表 3　記号を含まない形態素連続

/今/なら/送料/無料/
/パルック/蛍光灯/
/クール/色/

宣伝表現など商品の判別に寄与しないフレーズを**ノイズ**と定義する。一方で商品の判別に有効なフレーズを**シード**と定義する。例えば「今なら送料無料！」はノイズであり「パルック蛍光灯」や「クール色」はシードである。ノイズあるいはシードをフレーズの**ラベル**とよぶ。フレーズがノイズあるいはシードどちらとも言い切れない場合、便宜的に**中立フレーズ**とよぶ。ここで定義したノイズやシードなどのラベルは商品識別の際の重要度のみを考慮し、ビジネス的な重要度とは関係しない。

3.2 機械学習によるラベル判定

我々の目的は商品タイトルに含まれるノイズおよびシードを認識することである。これは商品データが与えられたときに商品タイトルに含まれる全フレーズに対してラベル判定を行うことで実現できる。判定は機械学習の二値分類によって行う。学習器には実装が容易な Voted Perceptron[2] を採用した。また問題を単純化するため、学習データおよびテストデータのうち人手で中立フレーズ判定されたものは事前に除外した。

3.3 ストア DF、カテゴリ DF

我々はフレーズのラベル判定を行う際にそのフレーズが出現する商品データのストア ID およびカテゴリ ID の頻度情報が利用できるのではないかと考えた。ある商品データ集合においてフレーズ f を含む全商品データの異なりストア ID 数をフレーズ f の**ストア DF** とよぶ。同様に異なりカテゴリ ID 数をフレーズ f の**カテゴリ DF** とよぶ。

ノイズはストアが商品を宣伝するために付与したものなのでストアに特有のフレーズが用いられることが多い。またストアは一般に複数のカテゴリに出品しているため、同一フレーズが複数のカテゴリで使われることが多い。従ってフレーズがノイズであればストア DF は低くカテゴリ DF は高いと予想される。一方でシードは商品特有のものなので複数のストアで共通して利用される。反面そのような商品は限られたカテゴリに偏ってあらわれる。従ってフレーズがシードであればストア DF は高くカテゴリ DF は低いと予想される。

予想を確認するため Yahoo!ショッピング[*1] 中の商品データに含まれるフレーズからノイズおよびシードを 200 件ずつ人手で収集して調査を行った。収集したフレーズのストア DF およびカテゴリ DF について対数をとり分布を調べた結果を図 1 に示す。この調査結果からノイズはストア DF に対してカテゴリ DF が高く、一方シードはカテゴリ DF に対してストア DF が高いことがわかる。なおここで用いたノイズおよびシードは各フレーズに対してそれぞれ 2 名の担当者が判定を

[*1] http://shopping.yahoo.co.jp

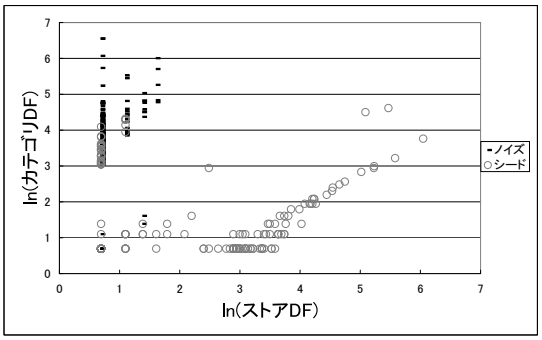

図 1　ストア DF、カテゴリ DF に関するフレーズの散布図

表 4　フレーズごとのストア DF・カテゴリ DF 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

行い両者の判定結果が一致したものだけを用いた。収集したフレーズの一部を表 4 に示す。

以上の考察からストア DF およびカテゴリ DF を機械学習の素性として用いることが有効だと考えられる。

4　精度評価

我々は 3 節の提案手法に基づいてラベル判定システムを開発し評価を行った。評価指標にはノイズおよびシードそれぞれのラベルについての精度 (precision) を用いた。

4.1　ラベル判定のガイドライン

3.3 節の調査において実際に人手で判定した総フレーズ数の 2 割程度は中立フレーズであった。そこで中立フレーズをカテゴライズし各々のカテゴリに属するフレーズに対してラベル判定の指針を与えるためのガイドラインを作成した。これを表 5 に示す。明確なガイドラインを策定できなかったカテゴリについてはラベルの欄に中立と記述した。これらの中立フレーズは今回は扱わないこととした。

表 5　ラベル判定のガイドライン

カテゴリ	ラベル	フレーズ例
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/
商品カテゴリ	ノイズ	/モダン/アジアン/家具/
ブランド	シード	/アル/ベロベロ/
社名 (製造)	シード	/日　立/
社名 (小売)	ノイズ	/浪花/商人/
社名 (製造・小売)	シード	/カワセ/
人名	ノイズ	/掛布/雅之/
形状・容量	シード	/ミニ/ボトル/５ｍｌ/
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

学習データを二種類用意してその違いによる精度比較を行った。

ひとつは Yahoo!ショッピング中に頻出するフレーズを人手で判定したものに加えてウィキペディア[*2] から自動収集したフレーズを加えたものである。これを学習データ A とする。こちらはデータ件数は多いものの品質は充分とはいえない。

もうひとつは学習データ A を用いて Yahoo!ショッピング中の全フレーズに対してラベル判定を行い、判定結果のうち分離平面に対して遠いほうから上位 500 件ずつを人手でラベル修正したものである。これを学習データ B とする。こちらはデータ件数は少ないが充分に精査したもののみ用いており品質はよい。

テストデータは Yahoo!ショッピングの商品データに含まれるフレーズを 300 件ランダムサンプリングして人手でラベル判定を行ったものを用いた。

機械学習の素性はカテゴリ DF に対するストア DF の比率のみを用いた。以降この素性を**フレーズ DF 比**とよぶことにする。なおストア DF およびカテゴリ DF は Yahoo!ショッピングから抽出した商品データ集合に対して算出した。各々のデータ件数を表 6 に示す。

評価結果を表 7 に示す。人手で精査した学習データ B を用いた結果シード判定精度が若干低下したもののノイズ判定精度は向上している。

[*2] http://ja.wikipedia.org

表 6　各種データ件数

データ名	件数 (ノイズ数)
学習データ A	2,428(598)
学習データ B	603(134)
テストデータ	286(99)
Yahoo!ショッピング 商品データ	5,848,419
Yahoo!ショッピング フレーズ	4,831,038

表 7　学習データ別評価

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

表 8　素性一覧

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

学習データに依らずシード判定精度が良好であり本システムを用いることで質のよいシードリストを自動作成できることが示された。一方ノイズ判定精度はシードに比べると低い。これは図 1 においてノイズの周辺にいくつかのシードが存在することからも納得できる。ノイズと誤判定されたシードには表 5 のガイドラインで例示したような、人が見ても判断に迷うフレーズが多かった。しかしノイズは絶対数が少ないためノイズ判定されたフレーズのうちスコアの高いものを改めて人手で精査することで高品質なノイズリストを作成することができる。

4.3　素性の違いによる評価

フレーズ DF 比とそれ以外の素性とを組み合わせ精度比較を行った。比較に用いた全素性を表 8 に示す。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** とよぶ。またフレーズ 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**とよぶ。評価結果を表 9 に示す。この結果からフレーズ DF 比を単独で用いた場合が最も精度がよいことが言える。

表 9　素性別評価

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

5　おわりに

本稿では商品タイトルに含まれるフレーズをラベル判定するための手法を提案し、高品質なシードリストが自動生成できることを示した。ノイズについては精度は充分ではないが判定結果を用いることで人手によるノイズリスト作成の効率化が期待できる。

今後の課題としてノイズ判定の精度をより向上させることが挙げられる。またガイドラインを策定できなかった中立フレーズの扱いについても検討する必要がある。

参考文献

[1] A. Aizawa and K. Oyama. A Fast Linkage Detection Scheme for Multi-Source Information Integration. *Web Information Retrieval and Integration, 2005. WIRI'05. Proceedings. International Workshop on Challenges in*, pp. 30–39, 2005.

[2] Y. Freund and R.E. Schapire. Large Margin Classification Using the Perceptron Algorithm. *Machine Learning*, Vol. 37, No. 3, pp. 277–296, 1999.

[3] M.L. Lee, H. Lu, T.W. Ling, and Y.T. Ko. Cleansing data for mining and warehousing. *Proceedings of the 10th International Conference on Database and Expert Systems Applications (DEXA)*, pp. 751–760, 1999.

[4] 山下達雄. 用語解説: Suffix array. 人工知能学会誌, 15(6):1142, 2000.