多目的クラスタリングを用いた情報推薦システム

千田 智治

1　はじめに

現在，多くの人が商品の購入を検討する際，既に商品を利用している人の声を参考に商品の購入を決定していると言われている．商品を購入する前にネット検索を行い，検討した後その商品を購入し，利用した感想を Blog や掲示板などで発信するという一連の流れを広告業界では AISAS の法則 [1] と呼び，現在の消費者の行動原理として利用されている．

商品の宣伝には作者やメーカー，広告代理店など「消費者に購入させたい」という思いから商品の情報が発信されており，実際に利用した消費者の口コミの方が信頼性は高いと思われる．実際に現在では，ただ広告を流しただけでは消費者は商品を購入しにくくなっており，ネットや友人からの口コミを重視して購入を決めている人が増えている．

商品の口コミは Amazon の Review を始め，価格.com など様々なサイトから発信されているが，消費者が購入する際にのみ閲覧されるものであり，消費者自らサイトを閲覧しにいくプッシュ型の行動が必要である．消費者という中立的な立場から集められた情報を有効活用したプル型の情報推薦システムがあれば，消費者にとって信頼のできる非常に有益な情報提供手段になると考えられる．

そこで本研究では，商品の Review を解析することで，消費者に信頼性の高い情報を発信し，類似した商品を推薦するシステムを提案する．

2　提案システム

2.1　概要

本システムでは，商品を消費者視点で分類し，その商品 Review を元に類似した商品を視覚的に推薦する．ユーザは予め分類されたクラスタを選択することで興味のある商品にたどり着ける．このシステムによって，商品の作者やメーカーなどから発信される情報ではなく，消費者という中立的な立場の視点から商品を推薦されることが期待できる．

次節以降に示す技術をマッシュアップすることでシステムを実装している．

2.2　Amazon Web Services(AWS)

Amazon 内の商品の Review を取得するために，本システムでは Amazon Web Services(AWS)[2] を利用している．AWS は，Amazon での商品の検索，データの表示，購入等をサポートする API のことであり，Amazon のほぼ全てのデータを利用することが可能である．

AWS には，HTTP 経由の XML または REST・SOAP を介してアクセスすることができる．また，AWS は XML ベースのプロトコルを用いているため，Fig. 1 に示すように Amazon から得られる情報は XML 形式で返される．

Fig.1　Amazon Web Services(出典：自作)

本システムでは，商品タイトル及び，ASIN，Review，画像，リンクなど特定の情報を取得する為，AWS で得られた XML から情報を抜き出す作業を行う必要がある．

2.3　形態素解析

Amazon から取得した商品の Review をオープンソースの MeCab[3] を用いて形態素解析を行い，その商品の重要キーワードを抜き出す．

形態素解析とは自然言語処理の技術の一種であり，自然言語で書かれたある文章を形態素（意味を持つ最小の単位）に分割し，それぞれの単語の品詞を判別する．

Fig.2　MeCab を用いた形態素解析 (出典：自作)

本システムでは，商品の特徴となりうる名詞と形容詞のみを抜き出して使用している（Fig. 2）．

2.4　ベクトル表現

Amazon の Review を各語の重みから構成されるベクトルとして表現するが，各語の重みは tf*idf 法を用いて重み付けを行う．

tf とは term frequency の略で，ある文書 d に単語 t が含まれる頻度を表す．また，idf とは inverted document frequency の略で，単語 t が文書集合 N に出現する頻度が少ない程，単語 t が文書 d を特徴付ける単語であるというものである．単語 t が出現する文書数を $df(t)$ とすると，式 (1) で表される．また，単語 t の文書 d における重み $w(t,d)$ として，tf と idf の異なる観点を組み合わせ

た式 (3) で表される手法が tf*idf 法である．tf が大きく，df が小さい (idf が大きい) と単語の重みは大きくなる [4]．

$$idf(t) = log\frac{N}{df(t)} \tag{1}$$

$$w(t,d) = tf(t,d) \times idf(t) \tag{2}$$

tf*idf 法で重み付けされた各語によって，式 (3) で，本システムで対象としている Web ページ（Review）をベクトル表現する．w_{ij} は文書 d_i の単語 t_j の重みである．M は全 Web ページ（全 Review）に出現する異なる語の数である．

$$d_i = (w_{i1}, w_{i2}, ...w_{iM})^T \qquad j = 1, ..., M \tag{3}$$

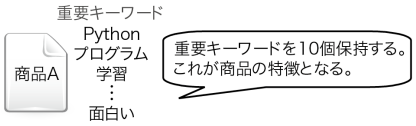

Fig.3　tf*idf 法を用いたベクトル表現 (出典：自作)

本システムでは，各 Web ページ (Review) が tf 値の高い単語 10 個を重要キーワードとして保持して（Fig. 3），2.5 節で示す類似度計算でそのデータを使用する．

2.5　類似度計算

2.4 節で求めたデータを元に，各 Web ページ（Review）がどのくらい類似しているかという指標である類似度を求める（Fig. 4）．

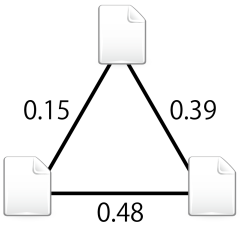

Fig.4　類似度計算 (出典：自作)

類似度の計算法は，Cosine-based Similarity, Correlation-based Similarity および Adjusted Cosine-based Similarity といった 3 つの方法があるが，本システムでは Cosine-based Similarity を用いる．Cosine-based Similarity では，2 つのアイテムは m 次元のユーザ空間のベクトルで表され，アイテム間の類似度は 2 つのベクトルが成す角度の余弦で計算される [5]．Web ページ（Review）が成す角度の余弦を式 (4) に示す．

$$S(d_i, d_h) = \frac{\sum_{j=1}^{M} w_{ij} w_{hj}}{\sqrt{\sum_{j=1}^{M} w_{ij}^2}\sqrt{\sum_{j=1}^{M} w_{hj}^2}} \qquad h = 1, ..., N \tag{4}$$

2.6　クラスタリング

2.5 節で求めた各 Web ページの類似度を元にクラスタリングを行う（Fig. 5）．クラスタリングには多目的シミュレーテッドアニーリングの一種である AMOSA をベースとした AMOSA クラスタリングを用いる．

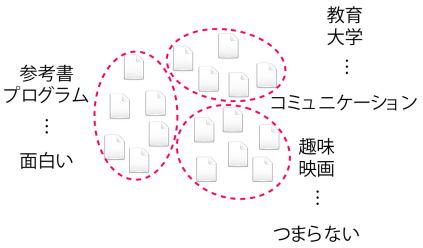

Fig.5　クラスタリング (出典：自作)

ノードを 1 つの Web ページ（1 商品の Review）とし，エッジを関連度の逆数とみなす．Web ページ間の類似度が高ければ類似度は大きくなるため，距離の概念を持つ AMOSA クラスタリングでは，類似度の逆数を距離として定義した．

AMOSA クラスタリングでは，分離性を表す評価関数である Overall Deviation と均質性を表す評価関数 Connectivity の 2 つの評価関数を同時に最適化することでクラスタリングを行なう．

Overall Deviation はクラスタの分離性に基づく大域的な評価指標である．Overall Deviation は各クラスタの中心点 μ_k から同じクラスタ内の各データ i までの距離の総和で定義される．この指標を最小化することにより，コンパクトなクラスタが生成される．この評価指標の数式を (5) 式に載せる．

$$Dev(C) = \sum_{C_k \in C} \sum_{i \in C_k} \delta(i, \mu_k) \tag{5}$$

Connectivity はクラスタの均質性に基づく局所的な評価指標として用いられる．Connectivity はあるデータ i の 1〜L 番目までの近傍 (近い順を j とする) が，データ i と異なるクラスタに存在する場合，$1/j$ のペナルティを与えるという指標であり，全てのデータについてのペナルティ値の総和が Connectivity 値となる．この指標を最小化することにより，近傍のデータ同士がより高い確率で同じクラスタに存在することになる．この評価指標の数式を (6) 式に載せる．

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} x_{i,nn_i(j)},$$
$$where\ x_{r,s} = \begin{cases} \frac{1}{j} & if\ \nexists C_k : r, s \in C_k \\ 0 & otherwize \end{cases} \tag{6}$$

AMOSA クラスタリングでは，任意に最大クラスタ数を決定することが可能であり，本システムでは最大クラ

スタ数を 25 として，クラスタ数 1 から 25 までのクラスタを生成させる．

2.6.1　解の生成

前々回の月例発表会では，AMOSA クラスタリングが NSGAII クラスタリングよりも精度が悪いことを報告した [6]．そこで，本システムを構築するにあたり，並行して AMOSA クラスタリングの精度向上を行った．

まず，探索において，現在の解から次状態の解を生成する際にどの部分に解を生成するのか調査を行った．調査の結果，クラスタ数 i の解から突然変異を行った場合，クラスタ数 i もしくは $i+1$ の解が生成されることがわかった．これは，突然変異の度に初期個体から近傍内の数本のエッジを結び直すため，クラスタ数 1 から 25 といった大幅にクラスタが変化しない為におこるものだと考えられる．

2.6.2　AMOSA クラスタリングの改良

AMOSA クラスタリングの改良にあたり，解の生成範囲をふまえ，クラスタ数 2 の解からクラスタ数 25 の解まで順に解の探索を行う手法を考案した．つまり，クラスタ数が少ない解からクラスタ数が多い解に向かって解を探索することで十分な探索を行っている．

AMOSA クラスタリング，NSGAII クラスタリング共に評価計算回数 24000 回として Fig. 6 に示すテストデータで評価実験を行った．テストデータはデータ数が 1000 の Square1 と Long1 を用いた．

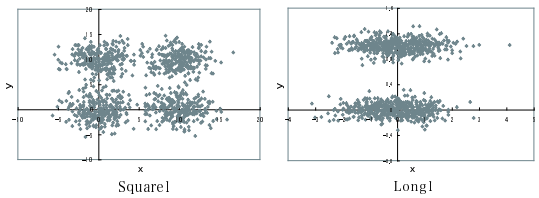

Fig.6　テストデータ (出典：自作)

Square1 のクラスタリング結果を Fig. 7 に示す．

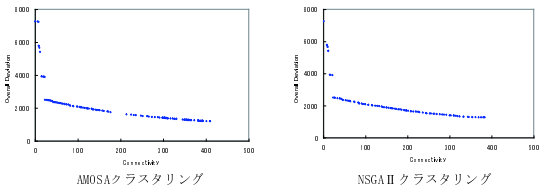

Fig.7　Square1(出典：自作)

また，Long1 のクラスタリング結果を Fig. 8 に示す．

Fig. 7 と Fig. 8 から AMOSA クラスタリングは NSGAII クラスタリングとほぼ同等の精度を示すことがわかった．よって，本システムではこの精度が向上した AMOSA クラスタリングを用いて，商品をクラスタリングした．

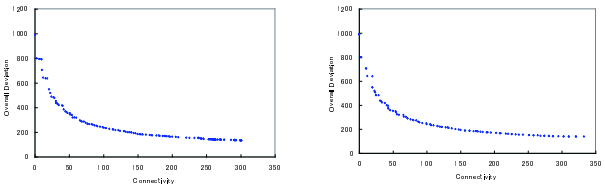

Fig.8　Long1(出典：自作)

2.7　SpringGraph

Amazon の商品 Review のクラスタリングした結果を，ばねモデルのライブラリ SpringGraph[7] を用いて UI を作成した．SpringGraph は Flex ベースのオープンソースライブラリである．

3　クラスタリング結果

クラスタリング結果を Fig. 9 に示す．Review を解析した結果をクラスタ数 1 から 25 まで表示しており，各クラスタを選択することでそのクラスタが保持する Amazon の商品データがページ遷移することなく表示される．

Fig.9　Review の解析結果 (出典：自作)

Amazon の商品データも類似度計算を行い，クラスタリングした結果で表示している（Fig. 10）．

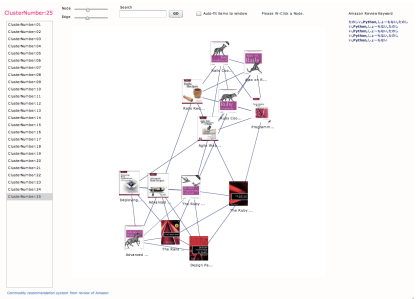

Fig.10　Amazon 商品の推薦 (出典：自作)

現段階では，本システムはクラスタリングした結果を表示するのみまでの進捗となり，そのクラスタリングが有効であるか，ベクトル表現で用いた重みは妥当なものなのか等，検討すべき課題がいくつか残っている．

4　まとめ

商品購入の際，消費者という中立的な立場で商品の分類が行われれば，信頼性の高い情報の推薦を行える．そこで，本研究では，Amazon Web Services から得られた商品の Review をクラスタリングすることで情報推薦型のシステムを構築した．

また，その際，クラスタリングには前回報告した AMOSA クラスタリングから精度を向上させたクラスタリング手法を考案し，適応させた．

本システムでは，クラスタリング結果の精度や，tf*idf を用いたベクトル表現の重み付けは妥当な値であるか等の検討が行えていない．よって，今後はユーザが探したい商品や興味のある商品を適切に推薦するできているか調査する必要がある．

参考文献

1) in clover LINE，AISAS（アイサス）の法則
   http://www.clover-line.jp/marketing/aisas.html
2) Amazon Web Services
   http://aws.amazon.com/
3) MeCab: Yet Another Part-of-Speech and Morphological Analyzer
   http://mecab.sourceforge.net/
4) 研究マップ自動生成システム, 大西 祥代，修士論文，2006.

5) 【文献調査】Item-Based Collaborative Filtering Recommendation Algorithms, 澁谷 翔吾, 廣安 知之, 三木 光範, ISDL Report No. 20081110001，2008.
   http://mikilab.doshisha.ac.jp/dia/research/report/2008/1110/001/report20081110001.html
6) 多目的シミュレーテッドアニーリングを用いたクラスタリング, 千田 智治，第 103 回月例発表会レジュメ，2008.

7) mark-shepherd.com
   http://mark-shepherd.com/blog/