テキストマイニングによるblogユーザの嗜好分析

テキストマイニングによるblogユーザの嗜好分析

澤井佑介[1)]，澤井政宏[2)]，久保　洋[1)]
1）室蘭工業大学情報工学科，　2）室蘭工業大学SVBL

1. はじめに

昨今，インターネットの普及にともない，blog と呼ばれるサービスが多く存在している．blog とは，Weblog の略でweb に残すlog という意味を持ち，インターネットに公開する個人の日記のことである．日記という体裁を取っているが，個人のプライベートな事が書かれるよりも，むしろ社会的に話題になっているニュースを引用し，批評やコメントを加えたりして，新しい視点を提供するといった方向性を持っているコンテンツである．その中にはある製品に対する感想やレビューが書かれている事があり，その製品を購入する動機付けになったり，その製品に対してユーザ同士の共感を呼ぶ事が出来る．

また，近年発達しているSNS（ソーシャル・ネットワーキング・サービス）も内部に「日記」と呼ばれるblog 的な機能を内包しており，広義的にはblog の一種であると言える．SNS には国内で著名なものとして，mixi やモバゲータウンなどが上げられる．

SNS はユーザ同士の繋がりというものを重視しており，多くのSNS では自分と同じ趣味・嗜好，同じ地域のユーザなどを探すための検索システムが提供されている．しかし，その検索システムはユーザ自身の登録情報に依存したシステムであり，またその趣味・嗜好などの登録情報は大まかな分類であり，詳細な嗜好の一致したユーザを探すことは困難である．また，登録情報と実際の趣味・嗜好が異なる場合があり，そのジャンルで指定して検索を行っても目的のユーザでない場合がある．

そこで本研究では，テキストマイニング技術を用いて，そのユーザの趣味・嗜好を分析し，コミュニティ形成の助けとなるシステムの構築を目的とする．

2. 用語辞書構築のためのデータ

本研究では，嗜好分析の前段階として，Amazon カスタマーレビューを利用し，用語辞書の構築を行う．

2. 1. Amazon カスタマーレビュー

Amazon カスタマーレビューとは，ユーザが自由に，感想・批評，レビューを投稿出来るAmazon.co.jp のサービスである．ユーザはレビューをAmazon に投稿し，それを見たユーザはそのレビューが商品購入に参考になったかの評価をする．

本研究では，ジャンル毎の売れ筋ランキングの上位120 位の商品のカスタマーレビューを収集し，そのデータを元に，用語辞書の構築を行った．

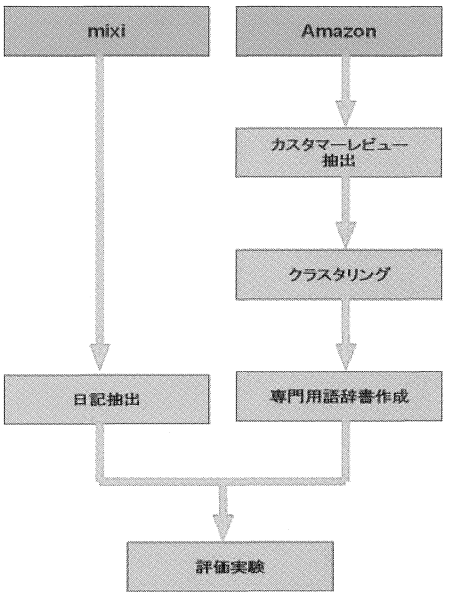

図1：実験全体の流れ

3. 嗜好分析

ユーザの嗜好分析は，ユーザがmixi 内で作成した「日記」に対して行われる．

その一例として，本研究では本のジャンル（例：歴史，ミステリー，ライトノベル等）に対するユーザの嗜好を分析する．嗜好分析のため，まずは「読書」の各ジャンルに特化した用語辞書の生成に重点を置く．

選択するジャンルはジャンルごとの差異が出やすいように，共通点の少ない組み合わせを5 個（ミステリー・時代小説・ライトノベル・技術書・経済経営）を選択した．そして，それぞれの分野で用語辞書を作成し，嗜好分析を行う．

3. 1. 用語辞書

用語辞書とは，テキストマイニングを行うときに用いる辞書であり，その分野で用いられる単語を集めた物である．用語辞書は手動で単語を入力し作成する方法もあるが，その方法では時間もかかり，製作者の主観も入るため内容に偏りが生じる恐れがある．そこで，本研究ではクラスタリング手法を用いることで用語辞書用の単語の選定を行う．2.1 節で収集したジャンル毎のカスタマーレビューに対して，茶筌を用いて形態素解析を行う．そして，それぞれのジャンルの単語の出現頻度を調べ，K-means 法でクラスタリングを行う[3]．

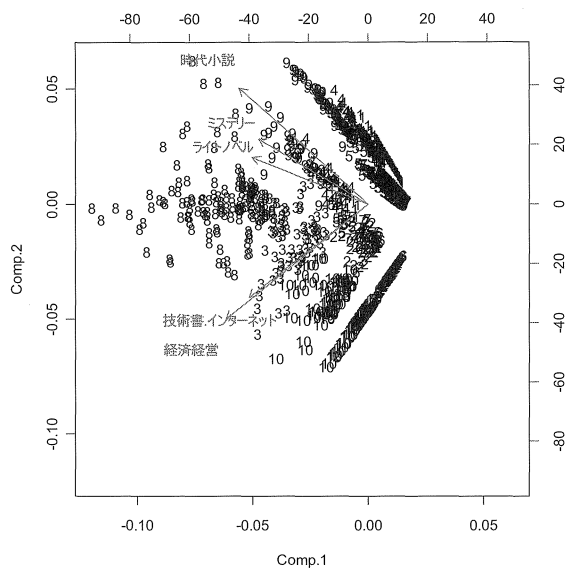

図2:K-means 法によるクラスタリング結果

これにより，単語をクラスタ毎に分類し，そのクラスタに登場する単語に基づいて，クラスタ毎に1～2種類のジャンルを割り当てた．実際にクラスタリングした結果の一部を表1および2に示す．

	ミステリー	時代小説	ライトノベル	技術書	経済経営
刑事	2.24	-3.76	-3.86	-3.78	-3.78
トリック	2.27	-3.76	-3.86	-3.78	-3.78
キリスト教	-0.76	-3.76	-3.86	-3.78	-3.78
最悪	-0.26	-3.76	-3.86	-3.78	-3.78
証拠	0.26	-3.76	-3.86	-3.78	-3.78

表1：ミステリーに割り当てたクラスタ(抜粋)

	ミステリー	時代小説	ライトノベル	技術書	経済経営
悲劇	0.33	-0.22	-3.86	-3.78	-3.78
英	-0.7	-0.82	-3.86	-3.78	-3.78
殺人	2.62	0.4	-3.86	-3.78	-3.78
リアリティ	0.7	-0.54	-3.86	-3.78	-3.78
南部	-0.02	0.23	-3.86	-3.78	-3.78

表2：ミステリーと時代小説の両方を割り当てたクラスタ(抜粋)

3. 2. 分析方法

出来上がった各ジャンル毎の用語辞書を用いて，日記の嗜好分析を行う．その方法として，まず日記に対して形態素解析を行い，単語毎に分解する．その単語を用語辞書と照らし合わせ，単一ジャンルが割り当てられたクラスタ内にその単語が存在すれば，1点を，2つのジャンルが割り当てられたクラスタに含まれていれば，0.5点をジャンル毎に加算した．日記に出現した全ての単語に対して上記の処理を繰り返すことにより日記を書いたユーザの嗜好度合いを算出した．

4. 評価結果と考察

4. 1. 評価結果

今回は4つのサンプルに対して，嗜好の度合いを算出した．そして，そのサンプルを被験者に実際に読んでもらい，そのサンプルがどの程度5つのジャンルに関係があるかを，アンケートにて5段階で評価させ，算出したサンプルの嗜好度合いとの相関を取った．

	相関係数
サンプル1	0.93
サンプル2	0.99
サンプル3	0.93
サンプル4	0.76

表3:評価結果

4. 2. 考察

今回の実験では推定されたサンプルの嗜好度合いとアンケートによる評価の相関がうまく取れており，用語辞書による嗜好分析はおおむね成功していると言える．しかし，サンプル4は他のサンプルと違い若干相関係数が低かった．

サンプル4に対する嗜好分析では，ミステリーに対して高い値が出ている．これは，ミステリーのクラスタに一般的に文章で使われる単語が比較的多く存在していたためであった．そのため，更なる辞書の精度の向上が求められる．

5. まとめ・今後の展望

本研究では，日記の内容からのユーザの嗜好分析を目的とした．実験結果から，用語辞書による嗜好分析が可能なことが明らかになった．

また，今回は時間の都合，あるユーザの一つの日記に対して嗜好分析を行ったが，今後はよりそのユーザの嗜好を詳細に把握するために，そのユーザが書いている複数の日記から趣味・嗜好分析を行う必要がある．

さらに，辞書にはまだ不要な単語が登録されていると考えられるので，更なる辞書の精度の向上が求められる．

本研究で，提案した手法によってSNSのコミュニケーションが活性化されることを期待する．

参考文献

[1]Amazon.co.jp： http://www.amazon.co.jp/

[2]松本裕治：形態素解析システム「茶筌」，情報処理 vol.41 (2000)

[3]K-means法：http://www.ie.osakafu-u.ac.jp/~honda/k-means.htm

[4]R on Windows：http://plaza.umin.ac.jp/~takeshou/R/