対話的なキーワード抽出によるブログ推薦システム

澁谷 翔吾

1　はじめに

近年，情報通信技術や情報社会の発展に伴い，社会が生成，管理する情報量が急速に増えている．このような状況の中で，ユーザにとって興味や価値のある情報を探し出すことは容易ではない．これらの背景から，情報の蓄積，検索，整理，およびアクセスなどにおいて，従来以上に効率的な手法や技術が求められている．そこで，近年では，ショッピングサイト，ニュース，映画紹介サイト，および音楽といった Web サービスにおいて，情報推薦といったユーザの情報収集を支援する手法が用いられている．これら情報推薦を行うには，いかにしてユーザの嗜好を抽出するかが重要となる [1)] ．

そこで，本研究では，システムが情報を推薦し，ユーザはその情報のどの部分に対して興味持っているか示すことを繰り返すことで，ユーザの嗜好を抽出する仕組みを考えた．的確な情報推薦を行うためには，システムが正しくユーザの興味を汲み取っているのか，ユーザのフィードバックが情報である．本提案では，情報を推薦する対象として，近年，急速に普及している Weblog(以下ブログ) を選択し，ユーザが興味あると思われるブログ記事を推薦するシステム（以下ブログ推薦システム）を構築した．

2　提案システム

本章では，提案システムについて解説する．提案システムはユーザが興味あると思われるブログ記事を推薦する推薦システムである．望まれる推薦システムとは，ユーザが要求してる情報を推薦するシステム，またはユーザが気付いていない興味に気付くようなシステムである．

2.1　概要

提案システムはユーザの興味に基づいてブログを推薦するシステムである．本システムでは，ユーザの興味はユーザにより示される仕組みとなっている．システムはユーザの興味，およびブログの特徴を蓄積し，個々のユーザの嗜好に合ったブログを推薦する．システムがユーザの興味を汲み取る仕組みを次節で解説する．

2.2　インタフェース

ユーザはシステムに初めてログインすると，ユーザプロファイルがないので，ランダムにブログのリストを提示する．ユーザはそれらのブログをクリックすることでブログ本文が表示され，自由に読むことができる．

ユーザにブログを推薦するためには，システムは何らかの方法でユーザの嗜好を得る必要がある．本システムでは，Fig. 1 に示すように，ユーザは興味ある単語，または文章をマウスでドラッグ操作（以下、ドラッグ）をすることで興味をシステムに伝える．ユーザはブログを自由に読みながら，興味ある単語，または文章をドラッグする．

Fig.1　インタフェース

ドラッグされた単語は，ユーザの興味を示す単語として蓄積され，文章がドラッグされた場合は，その文章を形態素解析し，名詞だけを抽出し，ユーザの興味を示す単語として蓄積する．形態素解析には，オープンソース形態素解析エンジン MeCab[2)] を用いている．システムにユーザの興味が蓄積されると，それを基にシステムはユーザにブログを推薦する．これらのブログも興味ある単語，および文章をドラッグを行うことでユーザの興味が蓄積される．

2.3　ユーザプロファイル

本節では，ドラッグにより示されたユーザの興味からユーザのプロファイルを作成する方法について解説する．

ドラッグによりユーザの興味がシステムに伝わると，その単語はそのユーザの興味ワードとして蓄積される．ドラッグ操作を繰り返すことで，Fig. 2 のようにユーザの興味ワードは蓄積されていく．

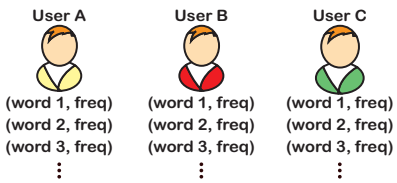

Fig.2　ユーザプロファイル

Fig. 2 における，word はユーザの興味ワード，freq はその興味ワードの出現回数を示している．本システムでは，この興味ワードとその出現回数を用いてユーザプロファイルを作成する．興味ワードの出現回数を考慮することで，そのユーザがその興味キーワードにどの程度，興味を示しているのかを捉えることができる．ユーザプロファイルを式で表すと以下のようになる．

$$P(UserA) = ((word1, frequency), (word2, frequency), \cdots, (wordN, frequency))$$

2.4 ブログの特徴

本節では，ブログの特徴付けについて解説する．ブログを特徴付けるには，文章中の特徴的な単語（重要とみなされる単語）を抽出するアルゴリズムである tf-idf[3] などが考えられるが，本システムでは，ユーザのドラッグ操作をブログの特徴としている．ユーザはブログを読み，興味ある単語，または文章をドラッグする．そのドラッグされた単語，および文章はユーザの興味ワードであると同時に，ブログを特徴付ける特徴ワードとしている．ユーザにより，特徴付けられたブログを Fig. 3 に示す．

Fig.3　ブログの特徴

Fig. 3 における，word は特徴ワード，freq はその特徴ワードの出現回数を示している．一つのブログは，多数のユーザにより特徴付けられ，ブログの特徴は以下のように表すことができる．

$F(BlogA) = ((word1, frequency), (word2, frequency), \cdots, (wordM, frequency))$

2.5 マッチング

ブログを推薦するためには，ユーザの興味を解析し，その興味にあうブログを選出する必要がある．本システムでは，ユーザの嗜好とブログの特徴が似ているものを推薦対象とする．ここで，似ているとはユーザの興味とブログの特徴との関連度が高いことをいう．

本システムでは，ベクトル空間法を用いる．ベクトル空間法とは，ユーザの興味とブログの特徴を多次元空間上のベクトルとして表現し，2 つのベクトルを比較することにより関連度を求める．ベクトルの方向はそれぞれの特徴を示すものであるので，2 つのベクトルのなす角が小さいほど似ている．これをユーザの嗜好とブログの特徴との関連度の計算に用いる．

上述したユーザプロファイル，およびブログの特徴付けはそれぞれ，単語とその頻度で表されることから，どちらもベクトルで表現することが可能である．つまり，ユーザプロファイルは N 個の興味キーワードとその頻度であることから，ユーザプロファイルは N 次元のベクトルと見なすことができ，ブログの特徴も同様に，M 個の特徴ワードとその頻度であることから，M 次元のベクトルと見なすことが可能である．このとき，頻度はその興味ワード，および特徴ワードに対する重みと考えられ，それぞれの総和が 1 となるように正規化を行う．これにより，ユーザプロファイルとブログの特徴との関連度は，以下の式で計算することができる．

$$sim(User, Blog) = cos\theta = \frac{\vec{User} \cdot \vec{Blog}}{\left|\vec{User}\right| \left|\vec{Blog}\right|} \quad (1)$$

シータはベクトル User とベクトル Blog のなす角であり，sim(User, Blog) は User と Blog の関連度である．

2.6 ブログの推薦

あるユーザと各ブログとの関連度を計算し，関連度の高いブログを推薦する．Fig. 4 の例では，ユーザ A と各ブログとの関連度を計算した結果，ブログ C が最も関連度が高いことからブログ C を推薦する．

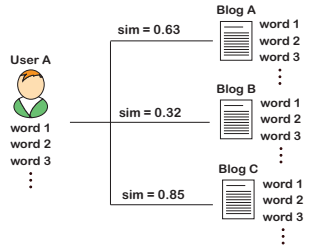

Fig.4　ブログの推薦

3　今後の課題

今後の課題として，2 点挙げられる．第一に，マッチングアルゴリズムの改良が挙げられる．現在のシステムでは，ユーザのドラッグにより得られた単語をユーザの興味，およびブログの特徴としているが，シソーラスを用いることで上位の概念を関連度計算に用いる．例えば，ユーザがゴールキーパーという単語に興味があった場合，そのユーザはゴールキーパーの上位の概念であるサッカーに興味を持っていると推測し，サッカーに関するブログを提示できる．上位概念で関連度を計算することで，マッチングの性能向上が期待できる．

現在のシステムでは，ユーザは興味ワードをリストで一覧できる．その中から特に興味あるキーワードを選択することで，そのキーワードを重視されるように改良する．キーワードに対する興味の度合いを変更することで，よりユーザの嗜好にあったブログが提示されることが期待できる．

4　まとめ

本研究では，ブログ推薦システムを提案した．本システムでは，ユーザの興味をドラッグの操作で行うことにより，システムがユーザの興味を汲み取る．この操作を行うことで，ユーザの興味を的確に捉えるシステムを目指す．ユーザプロファイルに基づいたブログの推薦には，ベクトル空間法を用い，ユーザプロファイルとブログの特徴との関連度を求め，関連度の高いブログを推薦する．今後の課題は，シソーラスを用いてキーワードを階層的に捉え，マッチングの精度向上，およびキーワードに対する興味の度合いを手動で変更できるように実装を行う．

参考文献

1) 土方 嘉徳, 情報推薦・情報フィルタリングのためのユーザプロファイリング技術, 人工知能学会論文誌 19 巻 3 号, 2004 年

2) MeCab
http://mecab.sourceforge.net/

3) tf・idf 法
http://www.forest.dnj.ynu.ac.jp/ ohmori/Paper/NL121/node6.html