分散化リコメンドシステムの提案

3L－9

橘高博行　佐藤直之　鈴木英明
NTTソフトウェア研究所

1. はじめに

現在、WWWで提供される情報を、ユーザの興味に応じて選択的に紹介する情報紹介サービスが数多く提案されている。しかし現状では、各サービスが独自にユーザの興味情報を収集して、ユーザプロファイルを設定している。したがって、ユーザが複数のサービスを利用する際には、それぞれのサービスごとにユーザプロファイルを設定することになる。このため、ユーザの興味が変化した場合には、すべてのサービスでユーザプロファイルを再登録する必要が生じる。また、ユーザの参照履歴等から自動的にユーザプロファイルを更新する場合でも、更新は各サービスごとに独自に行われており、他のサービスに自動的に反映されるわけではない。このように現状では、情報提供サービスとユーザとは1対Nの関係であり、円滑な情報紹介が行われているとはいいがたい。今後、情報提供者間でユーザプロファイルを共有し、N対Nの関係でシームレスに情報の紹介を行うことが求められている。

しかし、全サービスにおいて共通のユーザプロファイルを利用するためには、各サービスが提供する情報の全ての分野(項目)に対する興味度合いが、単一のユーザプロファイルで網羅されていなければならない。このとき、粗い概念で項目を羅列した場合は、項目の粒度の問題が生じる。例えば、「スポーツ」や「政治」のように範囲が広い項目を用いた場合は、スポーツの中で「野球」と「サッカー」のどちらにより興味があるか判断できない。逆に、単語レベルまで細分化された概念を項目とした場合は、ユーザの興味の全体像を掴みにくいという問題が生じる。この問題に対処するためには、類語辞書等を用いて、細かい単語レベルの概念同士を一つの大きな概念に関連づける作業を行い、大きな概念でユーザの興味を表現する必要がある。

本稿では、上記の問題を考慮し、個々のユーザが求める粒度で興味を表現することが可能で、なおかつ、類語辞書等を利用することなく興味の全体像が把握できるユーザプロファイルの表現方法を提案する。

2. 既存のユーザプロファイルの表現方法の問題点

既存のサービスでは、項目をどの粒度で扱うかで、大きく分けて2種類のユーザプロファイルの表現方法がある。

一つ目は、ある程度共通した意味解釈が可能な概念を項目に用いる方法である。この場合、(情報提供者が指定した項目，重み)のベクトルでユーザの興味が表現される。同様な方法で、情報がどのような興味を持ったユーザに適したものか、という情報の特性を表現する。例えば、ユーザプロファイルが((野球，10)，(サッカー，1))である場合は、「野球」により興味があることを示している。また、これと同じ重みが設定された情報は、「野球」に興味があるユーザに適した情報であることを示している。この方式では、意味解釈可能な概念をベクトルの項目としているため、ユーザの興味の全体像を把握し易い利点がある。しかし、これを全サービスで共通して用いる方法には、以下の問題がある。

・項目の意味解釈の統一が困難
　正しい情報選択を行うためには、項目の意味解釈を全サービスで統一する必要がある。複数のサービス間で、これを統一することは困難である。
・項目の管理が煩雑
　項目となる概念が重複して利用されていた場合には、情報を選択する際に不都合が生じる。新規に項目を追加する際に、既存の各項目と意味を考慮した照合を行い、重複して設定されることを防止する必要がある。

二つ目は、形態素解析等の手法を用いて各情報から抽出した単語を項目とする方法である。この方法は、1)情報に対してTF-IDF[1]等の統計計算を行い、それぞれの情報中の単語の頻度ベクトルを求める。ここで、各情報の特性は(単語，重み(頻度))で表現される。2)ユーザが参照した情報の頻度ベクトルを、ユーザプロファイルに反映させる。ここで、「ワールドカップ」や「Jリーグ」といった単語が高い頻度で出現する情報を多く参照したユーザは、これらの単語に対する重みが大きくなり、「サッカー」(という概念は意識されないが)に関連する情報が紹介されやすくなる。この方法を用いれば、項目を管理する必要がないため、前記の概念を項目として用いる方法の問題点を解決することができる。しかし、単語を項目とした方法には、以下の問題点がある。

・広い範囲の分野で提供される情報を扱った場合、ユーザの興味の全体像を把握できない
　ユーザが幅広い分野に対して同時に興味を持っている場合、満足のいく情報選択を行なえない可能性が高い。例えば、一般の新聞に代表されるような情報群から情報を参照した場合、一時「選挙」に関する情報を多く参照していると、「選挙」に関する情報ばかりが紹介され、「経済」や「スポーツ」に関する情報が紹介されにくくなる。

A Proposal of Dispersed Recommendation System
Hiroyuki KITTAKA, Naoyuki SATHO, Hideaki SUZUKI
NTT Software Laboratories

この問題は、複数分野にまたがるユーザの興味を単一のベクトル空間で表現していることに起因する。これを解決するには、情報の分野一つに対して一つの単語の頻度ベクトルを有するユーザプロファイルを用いる必要がある。本稿では、前記の方法を組み合わせ、単語の頻度ベクトルを基にしつつ、単語同士を関連づけた概念をベクトルの項目とするユーザプロファイルを作成する。

3. 提案するユーザプロファイルの表現方法の概要

本章では、前章で述べた問題を解決するユーザプロファイルの表現方法を提案する。類似する方法にWebMate[2]があるが、概念に重みを設定する点に本提案の特徴がある。提案方式の概要を図1に示す。

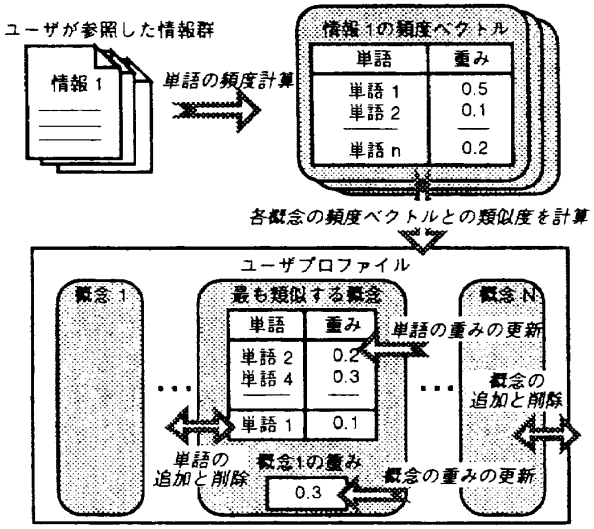

図1 ユーザプロファイルの作成

提案方法における、ユーザプロファイルの作成手順を以下に示す。

1) ユーザが参照した情報に対して形態素解析を行い、単語を抽出する。抽出された各単語の出現頻度から、各情報ごとに頻度ベクトルを求める。
2) ユーザの有する概念の個数がN個よりも少ない場合は、参照した情報の頻度ベクトルを、新しい概念を構成する頻度ベクトルとして追加する。
3) N個より多かった場合は、既存の各概念の頻度ベクトルと、参照した情報の頻度ベクトルとの類似度を計算し、最も類似する概念を決定する。
4) 3)の結果、ある一定値以上に類似する概念が存在しなかった場合、既存の概念の個数をN-1個にする。これは、既存の概念のうち最も重みの少ない概念を削除するか、類似する概念同士を一つの概念に縮退することで行う。そして、参照した情報の頻度ベクトルを、新しい概念を構成する頻度ベクトルとして追加する。
5) 新規概念の追加が行われなかった場合は、参照した情報の頻度ベクトルを、最も類似する概念の頻度ベクトルに加算する。加算することにより、すでに含まれていた単語の重みは増加し、存在しなかった単語は新規に追加される。同時に、概念の重みも増加させる。この重みは、ユーザの各概念に対する興味の度合いを示している。次に、頻度ベクトルにおいて、重みが大きい順に単語をソートする。頻度ベクトルに含まれる単語の総数がM個よりも多い場合、単語の総数がM個となるように、重みの少ない単語から順に削除する。

この方法は、ユーザが求める粒度でユーザプロファイルが設定されるという利点がある。例えば、広い範囲で「スポーツ」に興味があるのか「野球」に興味があるのか、ユーザの情報参照を通して自動的に設定される。「野球」に関する情報を多く参照すれば、自動的に「スポーツ」という大きな概念の中は「野球」に関する単語の頻度が高くなる。

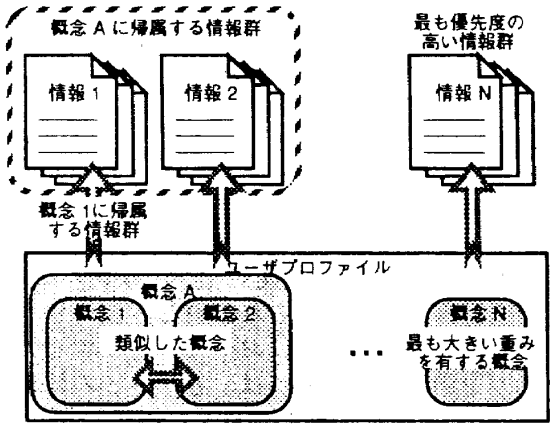

図2 情報の選択

情報の選択は、提供される各情報の頻度ベクトルと、各概念の頻度ベクトルの類似度を計算することで行う。類似度の高い情報から一定個数選択し、その概念に帰属した情報とする。これを図2に示す。この方法では、各概念ごとに情報を選択するため、ユーザが幅広い分野に対して同時に興味を持っている場合でも、各概念(分野)ごとに適した情報を紹介することができる。また、各概念の重みを考慮することで、ユーザの各概念に対する興味の度合いに応じて優先度をつけた情報紹介を行うことができる。また、類似した概念同士を一つの大きな概念として扱い、それに帰属する情報を同時に紹介することもできる。例えば、「野球」と「サッカー」に関する情報を一つにまとめて、「スポーツ」に関係する情報として紹介することも可能である。

4. おわりに

本稿では、複数の情報の分野が混在した環境においても、統一的にユーザの興味を表現することが可能なユーザプロファイルの表現方法を提案した。今後は、本提案方法を実装した情報提供サービスを運用し、有効性の確認を行っていく予定である。

参考文献

[1] Salton,G.:Developments in Automatic Text Retrieval,Science, Vol.253,pp.974-980(1991)
[2] Chen,L.:WebMate: A Personal Agent for Browsing and Searching, The Second International Conference on Autonomous Agents (Agents'98), (1998)