システムソフトウェアと
オペレーティング・システム 74－7
マルチメディア通信と分散処理 81－7
（1997. 2. 27）

キーワードを用いた情報フィルタリングシステムの構築に関する考察

藤川 英士　　窪薗 竜二　　保坂 憲一　　山森 和彦

NTTヒューマンインタフェース研究所
メディア応用システム研究部

情報がシステムに投入された際、その情報と利用者との適合度をスコアで算出し、スコアの高い情報のみを利用者に提供する情報フィルタリングシステムを構築中である。本報告の前半ではこのシステムを構築する際の課題および解決策として、プロファイルへの利用者の嗜好の表現方法、プロファイルと情報との適合度評価方法、プロファイルの修正方法について述べる。後半では提供した情報に対する利用者の行動から、その情報に対する興味の度合いを推測し、それをプロファイルの修正に活用する方法を検討する。検討材料として新聞記事を用いた実験を行い、結果から適合率を重視したシステムにおけるこの手法の有効性を明らかにすることができた。

A Study of Keyword-based Information Filtering System

FUJIKAWA Eiji　KUBOZONO Ryuji　HOSAKA Ken-ichi　YAMAMORI Kazuhiko

Multimedia Systems Laboratory
NTT Human Interface Laboratories

We have been developing Information Filtering System. It provides only appropriate information for the users by estimating adaptability of information. In the first half of this paper, we present the problems and its solutions —— how to express user's interest on the profile, how to evaluate an adaptability between it and information, and how to modify it. The latter half of this paper describes an experiment to investigate correlations between user's behavior toward information and user's favor. Experimental results show this method is useful for the system which aims to achieve high precision factor.

1 はじめに

通常、情報は利用者の検索行動とは独立してシステムに投入される。このため、たとえ利用者の興味を持っている情報が投入されても、それを察知できなければ、利用者はその情報を獲得することはできない。しかし、利用者自身がシステムに情報が投入されることを監視しておくことは多大な負担となる。このような情報獲得における利用者の負担を減らすしくみとして、情報フィルタリングシステムが注目されている。

情報フィルタリングシステムは、利用者はどのような情報が欲しいかをあらかじめ「ユーザプロファイル（以下、単にプロファイル）」に登録しておくことによって、利用者の嗜好にあった情報を提供する[1]。また、提供した情報に対する利用者の評価に基づいてプロファイルの更新または修正を行う。このとき

・利用者の興味の表現方法
・プロファイルと情報とのマッチング方法
・プロファイル修正・変更の基準

などが問題となり、これまで多くのアプローチが提案されている[2] - [6]。

この中で筆者らはテキスト情報がシステムに投入された際、その情報と利用者との適合度をスコアで算出し、スコアの高い情報のみを利用者に提供する情報フィルタリングシステムを構築中である。このシステムの主な特徴はプロファイルの修正・変更等の操作を利用者が直接行わなくても、システムに対する履歴情報によって、提供した情報に対する評価を推測し、その結果に基づいて自動的に処理を行うことである。

本報告ではシステムの構築の際に検討すべき課題を、利用者の行動の分析を中心に述べる。

2 情報フィルタリングシステムの構築

2.1 ベクトル空間モデルによる適合スコアの算出

テキスト情報を扱う個人適応型情報フィルタリングにおいては、ベクトル空間モデル[7]を用いた手法がよく知られている。この手法はテキストデータから文字列を検索するテキストサーチの技術が利用でき、処理速度と精度の点から有効である。

この手法によると、まず情報システム中に蓄積された全ての情報に含まれるキーワードにそれぞれベクトルを対応させてベクトル空間を定義する。この操作により、プロファイルを示すベクトルとフィルタリングの対象となる情報を示すベクトルが定義できる。これらのベクトルの距離を調べることによって利用者の嗜好に近い情報を獲得することができる。本報告では、利用者の嗜好と情報との近さをベクトルの内積を用いて算出し、これを適合スコアと呼ぶ。

2.2 プロファイルと修正ベクトル

プロファイルは、システムを利用する前に利用者に興味のある「分野」を答えてもらい、あらかじめ用意しておいたその分野に関するキーワード群のテンプレートをコピーすることによって作成する。

しかし、興味の対象をキーワードによって漏れなく記述することは難しい。そこで、上の方法で作成したプロファイルを初期状態とし、提供した情報に対する利用者の評価によってプロファイルを自動的に構築し、修正していく。

いま時点 t_n におけるプロファイルを P_n とし、利用者の評価を取得した時点 t_{n+1} のプロファイル P_{n+1} を考えると

$$P_{n+1} = P_n + \delta_n$$

となるあるベクトル δ_n が存在する。これを t_n における修正ベクトルと呼ぶ。t_n までのプロファイルがある程度正確に利用者の興味を表現していると仮定すると、修正ベクトルがなんらかの方法で算出できれば、より正確なプロファイルに修正することができると考えられる。この修正ベクトルの算出方法を得ることを本研究の最終的な目的とする。

2.3 システムの構成と処理手順

情報フィルタリングシステムとして、図2.1に示した構成のものを考える。

処理の手順は、情報がシステムに投入されたときの処理と、利用者の情報要求したときの処理は独立であるため、それぞれ個別に説明する。

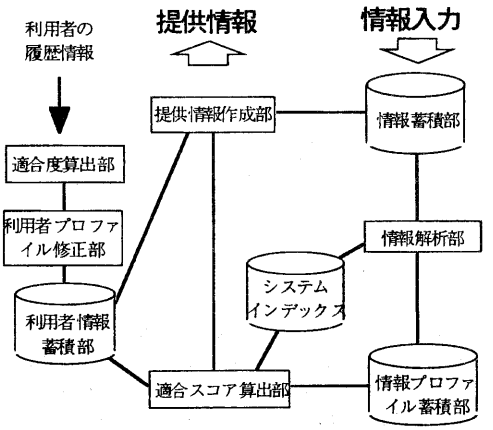


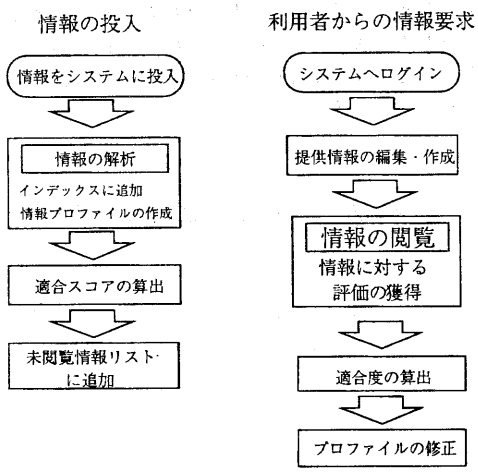
図2.1 システムの構成

2.3.1 情報が投入されたときの処理手順

図2.2(a)に情報が投入されたときのシステムの処理の手順を示す。

まず、図2.1の情報蓄積部に入力されたテキスト情報に対し情報解析部によってキーワード抽出処理を行い、キーワードとその出現回数から構成される情報プロファイルを生成する。同時にシステム全体の内容を表すシステムインデックスを修正する。システムインデックスはキーワードとそのキーワードの出現情報数をデータとしてもつ。

文献[8]に示されているように情報中に含まれる



(a)情報投入時　　(b)情報要求時

図2.2　システムの処理

キーワードの重みはシステムインデックス中のキーワードの出現情報数と、情報プロファイル中のキーワードの出現回数から算出する。これによりプロファイルと情報との適合スコアを算出する。算出した適合スコアは利用者ごとに用意した利用者情報蓄積装置中の未閲覧情報リストに保存する。

2.3.2　利用者の情報要求時の処理

利用者がシステムにアクセスした場合、図2.2(b)に示したように提供情報作成部において未閲覧情報リストに登録されている情報の中から適合スコアが高いものから順に、見やすい形に編集して利用者に提供する。

また、利用者がシステムからログアウトした時点で、提供した情報に対する利用者の評価に基づき修正ベクトル算出アルゴリズムによってプロファイルを修正する。

3　利用履歴からの興味の推測

3.1　利用者の行動の分析による興味の推測

情報を提供するたびに利用者にその評価を求めることは利用者にとって負担となる。そこで、提供された情報に対する利用者の行動を観察することによって利用者の評価を推測することを検討する。利用者の行動としては、情報の閲覧時間と嗜好とが深い関連性を持っていることが示されている[9]。さらに、提供した情報に対する関連情報の要求や検索行動、GUIにおけるウインドウの拡大縮小、スクロールの操作などが利用者の評価として利用できることが示されている[3][10]。

そこで、利用者の情報の閲覧時間と嗜好との関連性を定量的に評価できるという立場に立ち、閲覧時間から利用者の興味の度合いを推測する方法を検討する。

3.2　情報の閲覧時間と興味の有無の推測

例えば我々が新聞を読むとき、まず見出しを見て読む価値があるかどうかを判断する。読む価値があると判断すれば、時間を割いて本文を読み始めるし、読む価値がないと判断すれば次の記事に注意を移す。このことを逆に考えると、記事の閲覧時間によって、提供された記事に対する利用者の評価を推測できることが考えられる。

ここで検証しなければならない事項として以下のものをあげる。

(1) 見出しを見て行う読むか否かの判断はどの程度正確であるか

(2) 情報の閲覧時間とその情報に対する利用者の興味の有無はどの程度相関があるか

これらを検証するために次章以降で新聞記事を用いた実験について述べる。

4　要約情報による情報の適合度評価実験

4.1　実験の概要

この実験では3.2で挙げた課題(1)について検証を行う。新聞記事の場合、見出しは本情報である記事本文の内容や主題を簡潔に表した要約情報であると考えられる。扱う話題や利用者によって差があることが予想されが、要約情報やタイトルから判断された評価がどのくらい正しいかを調べる。

実験の手順は、まず被験者に一般紙から配信される比較的新しい記事の見出しのリストを見てもらい、記事の本文を読んでみたいかどうかを判断してもらう。さらに、同じ新聞記事を用いてそれらの本文を全て読んでもらい、「おもしろい」、「つまらない」のどちらかを評価してもらう。

これらのデータをもとに、見出しから判断した結果と本文を読んだときの評価の関係を調べる。5日間で約1000件の記事に目を通してもらう。

4.2 実験結果と考察

各記事に対して、見出しを見て本文を読むべきであると判断された記事を、検索された記事とみなす。また、実際に本文を読んだ結果、被験者に「おもしろい」と評価された記事を適合記事とする。このとき、各被験者についての適合率および再現率の分布は図4.1のようになった。

適合率 … 検索した結果得られた記事に含まれる適合記事の割合
再現率 … 全ての記事中の適合記事に対して、検索によって得られた適合記事の割合

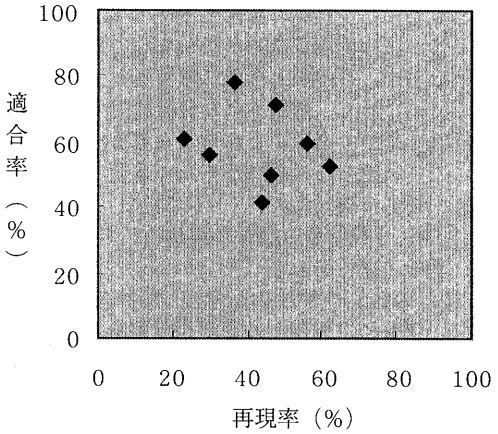


図4.1 システムの処理

さらに表4.1に各被験者について見出しの判断がどの程度正確であるかの割合で示した。表中(a)は見出しを見ておもしろいと判断し、かつ本文が適合していた率、(b)は見出しでつまらないと判断し、かつ本文が不適合であった率を示す。

この表によると、見出しを見たときの適合率、再現率はほぼ1/2たらずしかないことが分かる。しかし、見出しを見てつまらなそうであると判断したときに本文がつまらない割合は高く、見出しによって不適合記事と判断するときの正解率が高いことがわかる。

表4.1 見出しによる判断の正解率

被験者	適合記事数	(a)見○→本○	(b)見×→本×
a	88	45.0	87.8
b	53	53.1	92.6
c	91	49.3	86.2
d	148	28.7	81.4
e	134	40.8	79.4
f	70	47.9	89.0
g	167	43.5	75.2
h	75	37.0	90.0

5 情報の閲覧時間による興味の推測実験

5.1 実験内容

本実験では3.2の(2) に示した課題について検証を行う。以下、実験の内容を示す。

5.1.1 実験の手順

被験者に新聞記事を提供し、各記事について「おもしろい」、「つまらない」のどちらかを評価してもらう。その記事を評価し終わると次の記事を閲覧することができる。このとき、記事を提供してから評価が入力されるまでの時間を閲覧時間として計測し、利用者の評価と閲覧時間の関係を分布図に表す。

5.1.2 被験者

情報処理系の研究者13人を対象とする。被験者の職業は専門的であり、長期的に自分の携わっている分野に興味があり、その分野に関する有益な情報を獲得したいと思っていると考えられる。

5.1.3 記事の種類

工業新聞から比較的日付の新しい1000件の記事をおよそ１０日間に渡って提供した。工業新聞を選んだ理由は以下の通りである。

- 他の情報ソース（テレビなど）で公表されていない情報が多く、被験者にとって新しい情報が多い
- 狭い範囲に限ることによって利用者の記事の選択基準を明確にする
- 被験者が興味をもちそうな記事が適度にある

5.1.4 被験者への注意事項

記事の閲覧に関する注意事項として、被験者に以下のことをお願いした。

- 有益な情報を得ようとすること
- なるべく他の作業をせずに専念して読むこと
- 読む速度は自分の普段のペースでよい

5.2 実験結果に対する評価の方法

「おもしろい」と評価された新聞記事（以下、適合記事と呼ぶ）の閲覧時間に関する分布と、「つまらない」と評価された新聞記事（以下、不適合記事と呼ぶ）の閲覧時間に関する分布について調べる。

記事の分布に対して、ある時間を閾値として設定し、この時間より閲覧時間が少ない記事を捨て、閲覧時間が多い記事を得る。（この実験ではこの操作を検索と呼ぶ）その後、実際のデータと比べ、その正解率を適合率および再現率によって評価する。

5.3 実験結果と考察

利用者の記事の閲覧時間と記事に対する評価の関係を示す。図5.1はある1人の被験者についての閲覧時間の分布を示している。ここで、グラフの横軸は単位量あたりの閲覧時間（秒／100バイト）であり、縦軸は記事の数を示す。

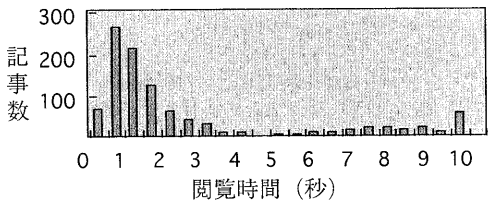


図5.1 閲覧時間に関する記事数の分布の例

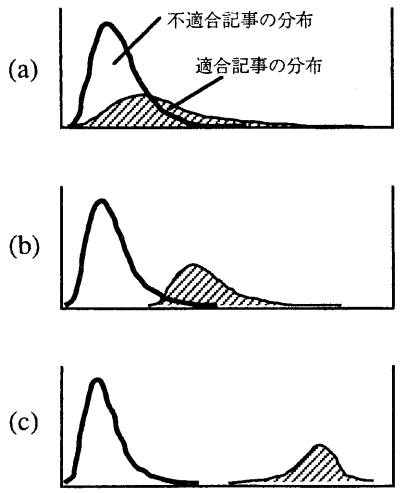


図5.2 閲覧時間に関する記事数の分布パターン

各被験者のグラフをみると、不適合記事のグラフについてはどれも似たものとなったが、適合記事のグラフは、被験者によっておおよそ図5.2に示すような3タイプの分布になった。この内訳は全被験者13人中、(a)3人、(b)7人、(c)3人となった。図5.1に示した被験者のデータは(c)に属する。

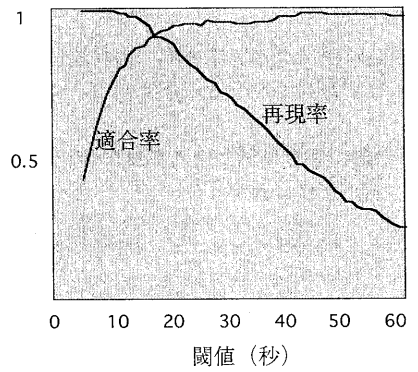


図5.3 閾値による適合率・再現率の変化

さらに閾値による適合率、再現率の変化を調べた。図5.3は図5.1の被験者のものである。

以上の結果から得られた考察を以下に述べる。

- 個人による結果の差が大きい。とくに図5.2(a)のような分布になる被験者は正解率が低い。
- 不適合記事の判断は記事の量について閲覧時間を正規化して評価しない方がよい。
- ほとんどの被験者について閾値を高くすると適合率は上がったことから、適合率を重視する情報提供方法、すなわち、大量の情報から重要な情報を少数とるようなシステムに適している。

6 情報の閲覧時間による興味の推測

6.1 情報の閲覧時間に関する仮定

前章までの実験結果と考察から、新聞記事の閲覧時間に関する仮定を述べる。

新聞記事においては、見出しを見て本文を読むかどうかを判断する。このとき、実験1により、見出しによってつまらない記事と判断したときの正解率が高い。すなわち、見出しを見たあと本文を読まない記事はおもしろくないものが多いことから、記事の長さによらず、ある閾値を決め、その値よりも小さい記事は削除すればよい。

削除後の分布において、単位量あたりの読む速度が一定と仮定すると、本文記事を読む時間はどこまで記事を読んだかを表す。このように考えると個人によって閲覧時間の分布に差がでる理由として被験者の読む速度の違いの他に、

・本文を全て読んだ後に記事の評価をくだす
・本文を読まずに記事の評価をくだす
・おもしろくないと判断した時点で本文を読むのをやめる

の3タイプに分けられることが推測される。また、閲覧時間による興味の有無の判断には、最後のタイプが適していることが分かる。

6.2 情報の閲覧時間による興味の度合いの推測アルゴリズム

前章までの実験から、システムが提供した情報に対する利用者の閲覧時間によって、その情報の興味の度合いを推測する方法を提案する。
以下、推測の手順を示す。

1．利用者が1回のアクセスにおいて目を通すことが可能な記事数nを決める。このときα（$0<\alpha<1$）を定数とし、n/α個の見出しを提供する。
2．利用者の閲覧時間および、閲覧した情報の量を保存しておく
3．閲覧時間がt_0よりも少ない記事は不適合として以後考慮しない
4．閲覧時間がt_0よりも多い記事の割合をNとする。
5．閲覧時間がt_0よりも多い記事について、記事の単位量あたりの閲覧時間を計算する
6．定数β（$0<\beta<1$）を用いて、単位量あたりの閲覧時間がβN番目に長いものまでを適合記事、短いものを不適合記事とする。

このアルゴリズムでは利用者が1回のアクセスにおいて閲覧する記事数が一定であることを前提としている。一定でない場合は、アクセスした際に、提供してほしい記事数としてnを利用者が決めるなどの方法をとる。

また、t_0および定数α、βは以下の意味をもつ。

・t_0　見出しをみて即座に判断するための所要時間
・α　情報提供の精度パラメータ
・β　見出しをみて「おもしろい」と判断したときに実際に本文がおもしろい割合

6 むすび

本報告の前半でキーワードを用いた情報フィルタリングシステムの構築に関してその課題と解決策を述べた。後半では課題を解決するための実験について述べた。今後は報告中に述べたように、内容の解析を行い、修正ベクトルの算出方法について検討していく。

謝辞

本報告の実験で被験者になっていただいたグループの方々に感謝致します。

参考文献

[1] Nicholas J. Belkin, and W. Bruce Croft, Information Filtering and Information Retrieval: Two Sides of the Same Coin? Comm. ACM, Vol. 35, No. 12, pp. 29-38, 1992
[2] 朝倉敬喜, 喜田弘司, 垂水浩幸, 宮下敏昭, エージェントによる情報フィルタリング, 情処研報, 情報メディア20-7, pp.49-55, 1995
[3] 神場知成, The Krakatoa Chronicle：WWW上のエージェント機能を利用した、対話型パーソナル新聞, 情処研報, システムソフトウェアとオペレーティング・システム, pp. 13-18, 1995
[4] 野美山浩, 紺谷精一, 渡辺日出雄, 串間和彦, 堤泰治郎, 個人適応型情報検索システム－個人の興味の学習する階層記憶モデルとその協調的フィルタリングへの適用－, 情処研報, 情報学基礎, pp1-8, 1996
[5] U. Shadanand, and P. Maes, Social Informtion Filtering : Algorithms for Automating "Word of Mouth", CHI '95
[6] D. Goldberg, D. Nichols, Brian M. Oki, and D. Terry, Using Collaborative Filtering to Weave an Information Tapestry, Comm. ACM, Vol. 35, No. 12, pp. 61-70, 1992
[7] G. Salton, Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer., Addison Wesley, 1989
[8] G Salton, Recent studies in automatic text analysis and document retrieval J.ACM, 20, 2, p.258, 1973
[9] 森田昌宏, 篠田陽一, 情報洪水の緩和のための情報フィルタリングの実現－ユーザアクティビティの分析と最適照合検索による情報フィルタリング, 第1回JAIN Consortium Symposium, pp.31-38, 1994
[10] P. Maes, Agents that Reduce Work and Information Overload, Comm. ACM, Vol. 37, No.7, pp.31-40, 1994