情報メディア 20－7
（1995．3．10）

エージェントによる情報フィルタリング

朝倉敬喜、喜田弘司、垂水浩幸、宮下敏昭
{asakura,kida,tarumi,miyasita}@obp.cl.nec.co.jp

NEC
関西C&C研究所

本稿では、エージェントベースの情報フィルタリングシステムについての提案を行う。対象となる情報は可読なテキスト情報とし、今回は特に電子メール・Internetニュースを対象にして説明する。フィルタリングでは情報の選別を行うだけでなく、その処理方法の決定と優先度の決定を行う。またフィルタリングにおいては興味の表現とその獲得が問題となるが、利用者の興味を短期的な興味、長期的な興味、さらに待ち行列に分け、それぞれにおいてフィルタリング判定を行う。本システムは我々が開発中のOA知的エージェントの一部であり、これはオフィスにおける生産性向上を目的としたものである。

Agent based Information Filtering

Takayoshi Asakura, Koji Kida, Hiroyuki Tarumi, Toshiaki Miyashita

NEC
Kansai C&C Research Lab.

In this paper, we propose an Agent based *Information Filtering* system. This system deals with text based information, especially electronic mail and Internet news. As well as selecting information, the system gives a priority value and ways of reaction to each piece of selected information. To acquire and to express user's interest properly, the system maintains three types of filtering parameters: short-term interest, long-term interest, and a waiting queue. This system is part of INA/LI(OA Intelligent Agent)system we are developing now, aiming at improving the office productivity.

1. はじめに

オフィスワークの生産性向上を計るために様々な試みがなされているが、OA化が急速に進み、これらの機器に縁がなかったユーザ層まで急速にその普及が進んでいる。

我々はネットワーク環境を前提とした総合的なオフィス業務支援を行うため、OA知的エージェント(INA/LI)の研究を行っている[1]。これはコンピュータ環境がある程度普及した電子化オフィスを対象とし、ホワイトカラーが行っている作業を代行することで、オフィス作業を可能な限り自動化し、その結果としてオフィスの生産性向上を計るものである。提供する各機能は部品化された複数の機能エージェントで提供され、核となる"ユーザエージェント"と通信しながら動作する。ユーザエージェントは利用者との対話、利用者の作業履歴の管理、各エージェントの制御を行うとともに、利用者の作業履歴から対話パターン等を獲得し、利用者に適応する秘書的インタフェースエージェントである。

これらの機能エージェントのうち、社内外のネットワークからの情報に対する処理(必要情報の選択や検索、他人への転送、不要情報の削除等)を支援するエージェントである情報フィルタリングエージェントに関する提案を本稿では行う。ここでエージェントとは、利用者に適応して代行処理を行うものであり、エージェント間の通信手段、利用者への適応手段、システムの状況を判定する手段を最低限保持するものである。

2. フィルタリングの目的

オフィスでは電子化された情報の活用は常識になりつつある。Internet、パソコン通信等のネットワーク上には様々な情報が蓄積されており、ローカルな機器上にも自分で作成した/ネットワークから到着した情報が保存されているが、情報量は膨大であり、必要とする情報がどれであるかの選択/検索や整理が必要となる。また、社内外問わず電子メールの発達によって個人宛に様々な情報が到着するが、一斉通知等の不必要な情報であることが多く、それら情報の選択に時間がかかることが多い。これらの問題を解決するために情報のフィルタリングが必要となる。

市販のツール等で実現されているフィルタリングは、メールのヘッダ等の定型的な情報を用いたフィルタリングであった。しかしながら、利用者が必要とする情報をヘッダ情報のみで表現することは難しいため、フィルタリング結果が利用者の求める情報と大きく異なってしまうことになる。そのため、利用者の興味を表現できるプロファイルを導入し、それを用いてある程度内容に踏み込んだフィルタリングを行うことが必要である。

ここで興味プロファイルは、

- 本来利用者自身が持っている興味
- 興味はないが仕事上必要な情報

から構成する必要がある。利用者の興味は利用者の入力した単語と利用者が情報に対して行った処理(履歴)から獲得し、仕事上必要な情報は他の業務エージェントと協調して獲得する。その結果、利用者の持つ興味や仕事に適応したフィルタリング、興味に基づいた情報の収集、作成等が可能になる。

例えば、フィルタリングエージェントとスケジューリングエージェントが連携(協調)した場合、以下の動作の(半)自動化が実現できる。

- 興味のある領域に関する催し等のスケジュールが到着した場合、スケジューリングエージェントを介して個人や部門のスケジュールに反映させる。
- 社内報告書の〆切や打ち合わせ等がスケジュールに入力された場合、利用者の興味に関係なくその領域に関するフィルタを新たにプロファイルに設定することで、情報収集を円滑に行う。

他のエージェントと連携した場合も、様々な効果が考えられるが、連携による誤った動作を防ぐために連携動作の指針をエージェント構築者があらかじめ明示しておく必要がある。また、他人のエージェントとの連携による動作(例えば、フィルタリング時に他人の興味に関する情報を参照して、他人が興味を持っている情報を転送する)も考えられるが、プライバシーの問題を解決する必要がある。これらについては、今後の課題とし、本稿では利用者の興味プロファイルを導入したフィルタリングエージェントの構成方法について提案する。

3. エージェントベース情報フィルタリング

3.1 対象情報媒体と処理内容

フィルタリングで対象とする情報は、メール、Internetニュース、ワープロの文書等の可読なテキスト情報とする。これらの情報をフィルタリングする場合、利用者の興味による採否判定(以下、単に判定と記す)を行うだけでなく、判定後の情報を提示するときに、今すぐ必要なのか、あとでまとめて参照するのかといった優先度を導入しソーティングするとともに、判定時に対象となった情報の処理方法も決定する必要がある。つまり、フィルタリング時には処理の決定と優先度の決定を行う。ここでの処理とは以下のことを示す。

- 情報を削除
- 他人へ転送
- 発信者へ返却
- ディレクトリへ保存
- 返事を書く

3.2 システム構成

図1にOA知的エージェントの全体システム構成を示す。本稿では特にフィルタリングエージェントについて述べるが、その他のエージェントについては[1]を参照していただきたい。

フィルタリングエージェントはエージェント動作の核となるエージェントエンジンを情報の入出力とし、単語抽出、フィルタリング、興味変換の各モジュールから構成される。

新規情報の到着時、あるいは利用者からの指示があった場合、単語抽出モジュールは後述する各データを用いてキーになる単語を抽出し、到着した情報に付加する。単語が抽出された情報の判定をフィルタリングモジュールが行い、その結果をメール・ニュースシステム、ユーザエージェントに返す。興味変換モジュールは短期興味を参照し、持続している短期興味を長期興味に変換する。以下、情報内容の表現、興味の表現と獲得、判定方法について説明する。

3.3 情報内容の表現

メール、ニュース等に存在するヘッダ情報は、フィルタリングの有効な判断基準となり得るが、情報の内容を表現するものとしては、Subject、Newsgroupしかない。さらに差出人等でもある程度その内容の推測は可能であるが、個人の興味を反映させたフィルタリングのためには不充分である。

そこで、テキスト情報から単語を抽出/選別し、それを情報の内容を示すタグにすることとした。単語抽出に必要なデータを以下に示す。

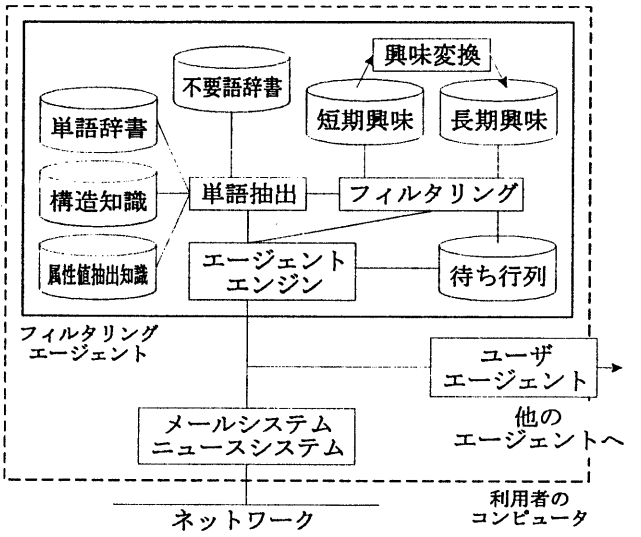

図1：システム構成

○単語辞書

単語抽出の基本となるデータで、抽出候補となる単語から構成する。単語は基本的に名詞であり、かな漢字変換辞書から学習結果を抽出して利用することにより、利用者に適応した単語の抽出を可能にする。

○不要語辞書

他の単語と複合した場合には意味をなすが、単独では内容表現に値しない単語、およびどちらの場合も内容表現に値しない辞書中の単語から構成する。

○構造知識

文章の題名、箇条書等は内容表現の上で重要なファクターであるため、情報の構造・書式から重要と考えられる部分を抽出するための知識を保持する。一例を以下に示す。

(BEFOREWORD=" 「 " , AFTERWORD=" 」 ") → WORD. Importance=1. 0

「」で挟まれた単語は重要である(単語抽出時における単語の重要度が最大である)、という意味である。

○属性値抽出知識

情報に含まれている属性名に対する属性値をフリーテキスト情報から抽出する知識である。フィルタリングに利用でき、抽出可能な属性は時間軸に関する属性(日付、時刻、〆切等)である。その他の属性については、他のエージェントが必要とする場合に知識を増強することで対応する。

属性名は、例えば属性名「日付」に対する属性値は「3月10日」等であり、「時刻」に対しては「15:00」等である。抽出された属性は、期限が過ぎた情報の消去、利用者への簡潔な情報提示に利用するが、スケジューリングエージェント等との連携にも利用する。知識の例を下記に示す。

(WORD="時刻") → SEARCHWORD(##:##)

(WORD="日付") → SEARCHWORD(#月#日, #/#)

これらのデータを用いて抽出された単語から内容を適切に表現していると推測される単語を以下の手順で選択する。

(1)属性値抽出知識によって抽出される単語

(2)単語辞書で抽出された単語の出現頻度をカウントし、そのカウントが高いものからある一定数以上の単語

(3)ヘッダ情報がある場合、属性名=属性値の型式に変換した単語

抽出/選択された単語群を情報の内容を表現している単語とし内容単語と呼ぶ。以下に一例を示す。各行が上記(1)(2)(3)に対応する。

日付 = 3月10日
KWD = エージェント、フィルタ
From = asakura@obp.cl.nec.co.jp

3.4 興味の表現とその獲得

利用者興味の表現とその獲得は、フィルタリングの一つの課題であるが、ここでは以下の方法を持って実現する。

(1)利用者の興味を短期的興味と長期的興味に分けて管理する。

オフィス業務においては、担当業務は長期的な(年単位)ものであり、その業務の中で様々な短期的な仕事(市場調査、他社動向調査、営業等)が発生すると考えられる。

一般に、利用者の持つ短期的な興味は、興味をもっている領域を表現する単語とその単語間の関係が明らかになっていないため、興味対象の空間は狭く、長期的な興味は、利用者の経験により興味対象の空間を表現する単語と単語間の関係が明らかになるため興味対象の空間は広くなる。また、短期的な仕事に関する情報の優先度は(期限があるため)高く、情報の広がりもあまりない。一方、長期的な業務に関する情報の優先度はさほど高くない分、情報の広がりがある。これに基づき、利用者の興味を2つのプロファイルで管理することで、より業務に即した情報のフィルタリングを行うことができる。

(2)短期的興味は情報に対する利用者の操作履歴から獲得し、長期的な興味は短期的興味から生成する。

短期興味は、内容単語とその単語が含まれた情報に対するフィルタリング判定頻度、利用者の情報に対する操作履歴等の統計的な情報で構成する。一定期間内に利用者に参照されない内容単語は興味が薄れたとして削除(忘却)し、利用者による既読頻度が高い内容単語は長期興味知識に変換する。

(3)短期的興味、長期的興味ともに利用者による設定を可能とする。

利用者の興味プロファイルは完全な自動獲得が望ましいが、現状のセンサー技術では不可能であると考えられる。また、フィルタリングによる到着・参照情報の収束(フィルタリングの固定化)、情報コントロール(フィルタリングの手法を知った上でフィルタを通過するようメッセージを作成し、意図的に到着情報のコントロールを行う)の可能性があるため、利用者による自分の興味プロファイルの参照・設定手段の提供が必要となる。

以降、各データの内容とその構築方法について述べる。

○短期興味

以下の項目から構成する。一単語あたりの処理内容に関するヒストグラムを図2に示す。1単語ごとに下記の内容を保持する。

- 内容単語
- フィルタリング判定頻度、通過頻度、破棄頻度
- 利用者の既読頻度、未読頻度
- 利用者の情報に対する処理(削除、返送、転送、保存、未処理)頻度
- 単語優先度I
- 情報到着時刻と情報を読んだ時刻(=既読時刻)のタイムラグ

単語優先度Iは以下の式で定義するが、到着～既読のタイムラグにより最大0.25の修正を加える。

$$0 \leqq I = \text{既読頻度} \ / \ \text{通過頻度} \leqq 1$$

短期興味に記述されている頻度情報は各内容単語の以下のような性質を示している。

- 未読数>既読数の内容単語は利用者にとって優先度が低いことを示す。
- 情報に対する処理の頻度の高低はその処理の実行を自動化できる度合いを示す。

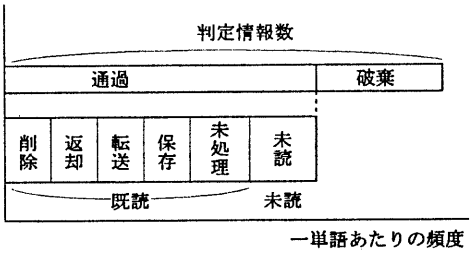

図2：一単語あたりの処理内容ヒストグラム

各頻度は、

- フィルタリング実行時
- 利用者による情報参照時、処理実行時
- 利用者による直接的な変更時

に更新し、その更新内容に基づき単語優先度Iを算出する。

○長期興味

以下の項目から構成する。

- 内容単語または内容単語で構成されたフィルタリング式
- 情報処理方法
- 利用頻度
- 最終利用日時
- 単語優先度I

長期興味は短期興味から生成することを基本としており以下の手順で生成する。最終利用日時が一定期間経過したものについては長期興味から削除する。

(1)短期興味における単語優先度の推移から長期間に渡って単語優先度が高い内容単語を検出する。

(2)検出された内容単語の処理頻度の高い処理を採用する。ただし、処理頻度に優位性が見られない場合はユーザにその決定を委ねる。ここでの優位性とは全頻度における各頻度の割合と定義する。

(3)単語優先度Iは0.5として長期興味知識に登録する。ただし、処理が削除の場合は付加しない。

フィルタリングモジュールは、フィルタ判定時に真となった内容単語の単語優先度Iを高める。偽となったものに対する単語優先度Iを低める。

○待ち行列

自分が発信した情報に対する返事を待つ場合が多々ある。このときの返事はあらかじめ「誰から」どのような返事・情報が到着するのか判明している場合、一時的にその情報の優先度を高く(あるいはフィルタの目を粗く)することで、それらの情報の獲得を容易にする。情報送信時の情報内容から返事が予想される場合、待ち行列にその情報の内容単語とヘッダ情報を登録する。返事が予測されるかどうかの判定は、内容単語中の特定単語(返事、お願い等)の存在の有無で行う。

返信されると予測した情報が到着した時点で、それに対応する待ち行列の内容は削除する。一定時間を経過した内容についても削除を行う。

3.5 判定方法

判定に用いるデータの形式が違うため、データごとに判定方法を説明する。判定が相違した場合は基本的に、待ち行列を最優先し、次に短期興味を優先して判定する。情報の到着から、単語抽出、フィルタリング(判定)、利用者による情報の参照までの流れを図4に示す。フィルタリングエージェントで行うのは点線内のみであり、それ以外はメール・ニュースシステムの動作である。

○待ち行列

到着情報が待ち行列登録内容に対する返事かどうかの判定は以下の通りである。

- 発信情報の「To」属性値が到着情報の「From」属性値であった場合
- 発信情報の引用が確認された場合

これが満たされた場合は、情報優先度IIを最大にして、フィルタ通過とする。

○短期興味

短期興味に保持されている内容単語から到着情報に含まれている内容単語に合致するものを抽出し処理頻度(削除、返却、転送、保存、返事、未処理、未読)を合計して処理Mを決定する。

未読と削除が最大の場合は破棄(情報優先度II=0とする)、転送の場合は過去の転送先へ転送、保存の場合は特定ディレクトリへ保存する処理とする。それ以外の場合は情報優先度IIを計算して通過とする。情報優先度IIは各単語優先度Iの平均値とする。

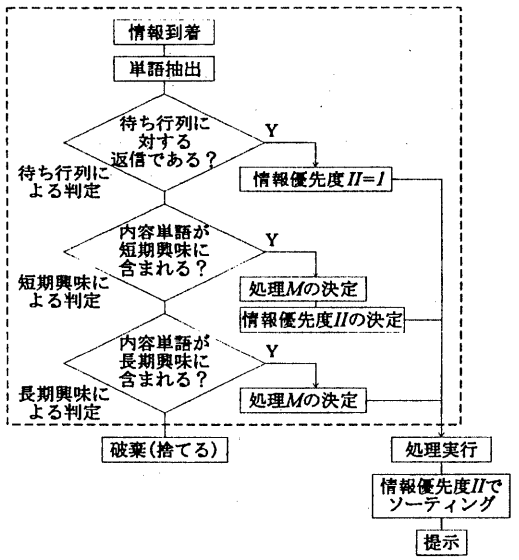

図4:情報到着から提示までの流れ

○長期興味

長期興味に保持されている内容単語から到着情報に含まれている内容単語に合致するものを抽出し、その情報処理方法に記述されている処理を処理Mとする。処理Mが相反する場合は削除しない方向で処理を決定する。ここでの削除処理はフィルタを通過しないことである。

○判定後

処理Mが決定された到着情報はメール・ニュースシステムまたはユーザエージェントに返される。メール・ニュースシステムは処理Mを実行するとともに情報優先度IIが高い(0.7以上)場合は利用者に到着メッセージを出す。中程度の場合(0.3-0.7)は到着したことをアイコンの変化等で通知し、低い場合は通知を行わない。

4. 考察

4.1 他システムとの比較

エージェントベースのフィルタリングシステムとしてはNewt[2]、TAPESTRY[3][4]、GroupLens[5]などがある。

Newtではユーザからのフィードバックのほかに他のエージェントと協調することで利用者の興味獲得を行い、利用者に適応した情報(Internet

ニュース等)フィルタリングを実現している。各エージェントごとに利用者適応を行い、適応知識を分散させている。また、エージェントのビジュアル化にも重点がおかれている。またフィルタリングの判定に関しては、情報とプロファイルの類似度を定義し、類似度の大小で判定を行っている。

TAPESTRYでは、ヘッダ情報を用いたフィルタリングに加え、同一の目的を持つ複数の利用者おのおのが、情報に対して「好き」、「嫌い」の意見を付加することで協調的なフィルタリング行っている。協調的なフィルタリング時には、個々の利用者の個人的な情報が参照される。GroupLensも同様のアプローチである。

本システムは上記システムに対して、以下の特徴を有する。

- 興味の獲得
  情報参照時に利用者には特別な入力を要求をせず、利用者に負担をかけずに興味の獲得を行う。
- 興味の保持方式
  興味に関する適応知識はすべてフィルタリングエージェントで管理し、それ以外の、例えば利用者との対話における適応知識(操作のくせ、対話パターン等)はユーザエージェントが管理することで、Newtで発生すると考えられる知識の重複を避けている。
- 計算量の抑制
  前記両システムでは情報の領域が広がった場合、フィルタリング判定時の計算量の増大が考えられるが、本システムでは負の興味情報をプロファイルに入れ、さらに優先度を定義し、それに基づいてソーティングを実行し、それ以降の選択(読む/読まない)を利用者に委ねることで計算量を抑えている。

TAPESTRYで行われている複数利用者によるフィルタリングは、プライバシーの問題も含めて今後の課題である。

4.2 単語抽出

内容単語の選択において、現在は頻度的に連続するすべての単語を採用している。表層的な記号処理では重要な単語を選択することができないためであるが、内容とは関連のない単語が常に混入する。これをいかにしてユーザのフィードバックで削除するか今後検討を要する。また、不要な上位概念を削除するためにシソーラスの導入を検討する必要がある。

4.3 判定方法

他のエージェントが同一マシン上で動作していることもあり、極力計算量を落とす必要があるため、行列演算等をせず、簡単な数値計算で処理と優先度を決定している。計算量を増やさず、より正確なフィルタリングを実現する判定方法を検討する必要がある。

5. おわりに

本稿ではエージェントベースでフィルタリングを実現するシステムについて述べた。本システムはOA知的エージェント(INA/LI)の1エージェントとして実装中である。

情報フィルタリングに続くオフィスにおける情報を核とした支援は、情報の要約、創造性であると考えられる。これら支援における課題を抽出して研究を進めてゆく。

参考文献

[1] 石黒他, オフィス生産性向上のための知的インターフェース, 第10回ヒューマンインタフェース・シンポジウム論文集, pp.143-146, 1994

[2] P.Maes, Agents that Reduce Work and Information Overload, Comm.ACM, Vol.37, No.7, pp.31-40, 1994

[3] Douglas B. Terry, et al., A Tour Through Tapestry, Proc. of COOCS'93, ACM No.11, pp.21-30, 1993

[4]D.Goldberg, Using Collaborative Filtering to Weave an Information TAPESTRY, Comm.ACM, Vol.35, No.12, pp.61-70, 1992

[5]P.Resnick, et al., GroupLens:An Open Architecture for Collaborative Filtering of Netnews, Proc.of CSCW'94, pp.175-186, 1994