Res. Rep. Fac. Eng. Mie Univ., Vol. 23, pp. 125−131 (1998)

総　説

情報洪水と情報フィルタリング技術について
Information Flooding and Information Filtering Technique

北　英彦
Hidehiko Kita
（電気電子工学科　Department of Electrical and Electronic Engineering）

(Received ...)

Abstract

A massive information and data is published on the internet through the medium of the WWW(World Wide Web) and so on. The user of the internet is able to get all sort of information and data published on the internet easily. However, it is difficult to pick up only the information and data which is useful to the user from the massive information and data. This difficulty is called information flooding. The information filtering technique is proposed and researched in order to resolve this difficulty. In this paper, a summary of the information flooding and the information filtering technique.

1．はじめに

インターネットの急速な普及に伴い，インターネット上にはWWW（World Wide Web）をはじめとする情報提供／情報入手方法により膨大な情報が提供されている．その結果，インターネットの利用者は，種々の情報を簡単に入手することができ，また，膨大な情報の中には利用者にとって有用な情報が含まれている潜在的可能性が高まっている．しかしながら一方，これらの情報の中から自分にとって有用である情報のみを効率的に取り出すことが難しくなっているという問題が発生している．この問題は情報洪水という言葉でも呼ばれる．この問題を解決あるいは軽減するための方法として情報フィルタリングという技術が提案され種々の研究がなされている．本総説では，最初にインターネットの現状について述べ，次に情報フィルタリング技術についての概観を述べる．

2．インターネットの現状

2．1　利用者数の現状

1960年代末に米国で誕生したインターネットはその後全世界に広がり，パーソナルコンピュータ（いわゆるパソコン）やモデムの低価格化，パソコン通信サービスとインターネットの相互接続，インターネット接続サービスの一般化などの種々の要因によって，一般社会へ急速へ普及しつつあ

る．

日本でのインターネット利用者は，「インターネット白書'97」[1]および「インターネット白書'98」[2]によると，1997年6月の時点で約570万人と推定されていたものが，1998年6月の時点では約2倍の約1000万人と推定されている．これは日本国民の10人にひとりがインターネットを利用していることを示している．家庭電化製品においては普及率が世帯数の30%程度を越えるとどの家庭でも利用していると感じられるようになるといわれているが，インターネットの利用者数も近いうちにその値を越えるものと予想される．また，従来はインターネットの利用は大学を主とする学校および企業を主とする勤務先からがほとんどだったのに対して，1998年の推定では，約190万世帯，約250万人，すなわち，インターネット利用者の1/4は，家庭からのみインターネットを利用していると推定されている[2]．これは，仕事のためのインターネットの利用だけではなく，趣味，娯楽，買い物など普通の日常生活の中でもインターネットが利用されはじめていることを表している．

全世界についても同様な状況であり，インターネット利用者のよい推定値がないためインターネットに接続されている計算機（ホストと呼ぶ）の数で述べると，1997年1月の時点で約1600万台がインターネットに接続されていると推定されていたのが，1998年1月の時点でこれも約2倍の約3000万台が接続されていると推定されている[1][2]．

2．2 インターネット上の情報量の現状

インターネット上の情報公開／情報入手の方法として幅広く用いられており，そのためインターネットそのものと間違って混同されることの多い，WWW（World Wide Web）の現状について次に述べる．「インターネット白書'97」[1]および「インターネット白書'98」[2]によると，WWWの情報提供の手段であるWWWページを提供しているサーバーの数は，1996年1月の時点で約10万台，1997年1月の時点で約65万台，1998年3月の時点で約208万台と推定されている．これもインターネットの利用者数と同様に，1年間でほぼ2倍の増加率を示している．WWWページの数についてはよい推定値がないため，検索ページ（世界中にあるWWWページを登録しておいてキーワードによってWWWページを検索できるようにするシステム）として有名なAltavista[3]に登録されているWWWページ数で見てみると，1997年5月の時点で約5千万ページ，1998年5月の時点で約1億1千万ページとなっている．これも，他と同様に1年間でほぼ2倍の増加率を示している．

次にインターネット上の情報公開／情報入手の方法としてインターネットの初期の時代から利用されているネットニュース（電子ニュースともいう）の現場について述べる．ネットニュース上の情報のひとつひとつは記事と呼ばれる文字情報である．記事は何に関する情報かということを表すために，予め準備されたニュースグループという分類を示してインターネット上に投稿される．ニュースグループは最初に決められたままで変更をしないというわけではなく，必要に応じて，あるいは，ネットニュースの運営者の判断によって，新規に作られたり，あるいは，削除されたりする．

ニュースグループの数は，1998年春の時点で全世界で1万以上，日本を中心としてやりとりされるfj（from Japanの意味）で始まるニュースグループが400以上存在する．ネットニュースに流される記事は，全世界で1日平均約40万通，fjで始めるニュースグループに限っても1日平均約2千通である．また，1日当たり100通以上の記事が投稿されるニュースグループも多数存在する．ネットニュースのごく初期の時点では，ネットニュースに流れる全ての記事に目を通すことが可能で著者をはじめ多くの利用者がそのようにしていたが，このように多量の記事が投稿されるようになると，かなりニュースグループおよび記事を絞り込んで記事を読むようするか，読むのをあきらめるしかないようになっている．

以上のデータから，ここ1年ではインターネットの利用およびインターネット上の情報は1年間に約2倍の増加率を示していることが分かる．今後は，普及度が進むことにより増加率は低下するもの

と予測されるが，米国を除く日本を含めた世界の他の地域ではインターネットは飽和するほどには普及していないので，インターネットの利用およびインターネット上の情報の絶対量が増えていくことには間違いがないと思われる．

2．3　インターネット上の情報の特徴

次にインターネット上でやりとりされている情報の特徴について考察する．インターネット上の情報の大きな特徴のひとつは，利用者が手軽に情報を入手できることである．インターネット上の情報公開／情報入手の手段の代表であるWWW（World Wide Web）に関しては特に顕著で，アンカーと呼ばれるリンク情報が割り当てられているアイコン（図記号）や文字列をマウスを用いてクリック（選択）するだけでリンク先のWWWページが表示される仕組みになっているため，知りたい情報をマウスで選択していくだけでその情報を得ることができる．ネットニュースに関してもニュースサーバへの接続の設定ができていれば，まずニュースグループを選択し，次にそのニュースグループに投稿された記事を選択するという方法で情報を入手することができるため，簡単な使い方さえ覚えれば簡単に情報を入手することができる．

インターネット上の情報のもうひとつの大きな特徴は，新聞，テレビ，雑誌などの既存の情報メディアとは異なり，誰でも簡単に情報を発信できることにある．WWW（World Wide Web）に関しては，HTML（Hyper Text Markup Language）と呼ばれる他の文書へのリンクを埋め込むことのできる記述言語を覚えさえすればWWWページの記述ができる．HTMLの持つ機能は文書の表示に関する直感的に分かりやすいものが多いため，プログラムを書くときのような計算機に関する専門的な知識を要求されない．多くのインターネット接続業者（インターネットプロバイダと呼ばれる）が提供している，WWWサーバーへの利用者のWWWページの登録のサービスを利用することにより利用者が作成したWWWページを全世界へ公開することが可能である．ネットニュースに関しては，テキスト（文字列）で記事を記述して，ネットニュースリーダーの投稿の機能を用いれば，その記事は全世界へ配信されていく．

このように誰でもが手軽に情報を公開することができるために，インターネット上の情報はさらに次のような特徴を持つ．

（1）誰でも手軽に情報を公開できるため，前節でも述べたように膨大な量の情報がインターネット上に溢れている．この事実のことを，情報洪水と呼ぶことがある．

（2）WWWサーバーは，計算機，ソフトウェアの管理に関して多少詳しく，また，管理している計算機をインターネットに常時接続することが可能ならば，比較的簡単に立ち上げることができる．このため，前節でも述べたように多数のWWWサーバーが情報提供のサービスを行っている．これらのWWWサーバーはそれぞれ関係なく独立して運用されているため，どこにどのようなWWWページを公開しているWWWサーバーがあるかを組織的に知る方法が存在しない．そのため，情報はインターネット上に無秩序に存在している．この問題を解決または軽減するために，検索ページ（世界中にあるWWWページを登録しておいてキーワードによってWWWページを検索できるようにするシステム）やリンクリストページ（ある話題／事物に関係するWWWページの一覧のWWWページ）が自然発生的に発展してきたものと考えられる．

（3）誰でも情報を公開できるために，情報の品質に保証がないというのもインターネット上の情報の特徴のひとつである．そのために，誰にでも役に立つWWWページや，専門家にとって役に立つWWWページが多数存在する一方，公序良俗や社会的なマナーに反するWWWページも多数存在する．

以上の特徴は，「品質の保証のない，膨大な情報が無秩序に存在する」というように要約することができる．

3．情報フィルタリング技術

「品質の保証のない，膨大な情報が無秩序に存在する」という情報洪水という問題に対して，その問題を解決あるいは軽減するひとつの方法として，情報フィルタリングという考え方が提案されている．

3．1　情報フィルタリングの一般的モデル

情報フィルタリングの言葉通りの意味は，「情報源の中から情報を選別すること」であるが，通常は情報洪水に対処するための技術として，

『膨大な規模の』情報源の中から
『有用な』情報を
『その情報が有用であることを自動的に推測して』
『自動的に』選別すること

として考えられている．この定義の意味を説明するために，また，情報フィルタリングに関する考察を行うために，情報フィルタリングの一般的モデルを図1に示す．

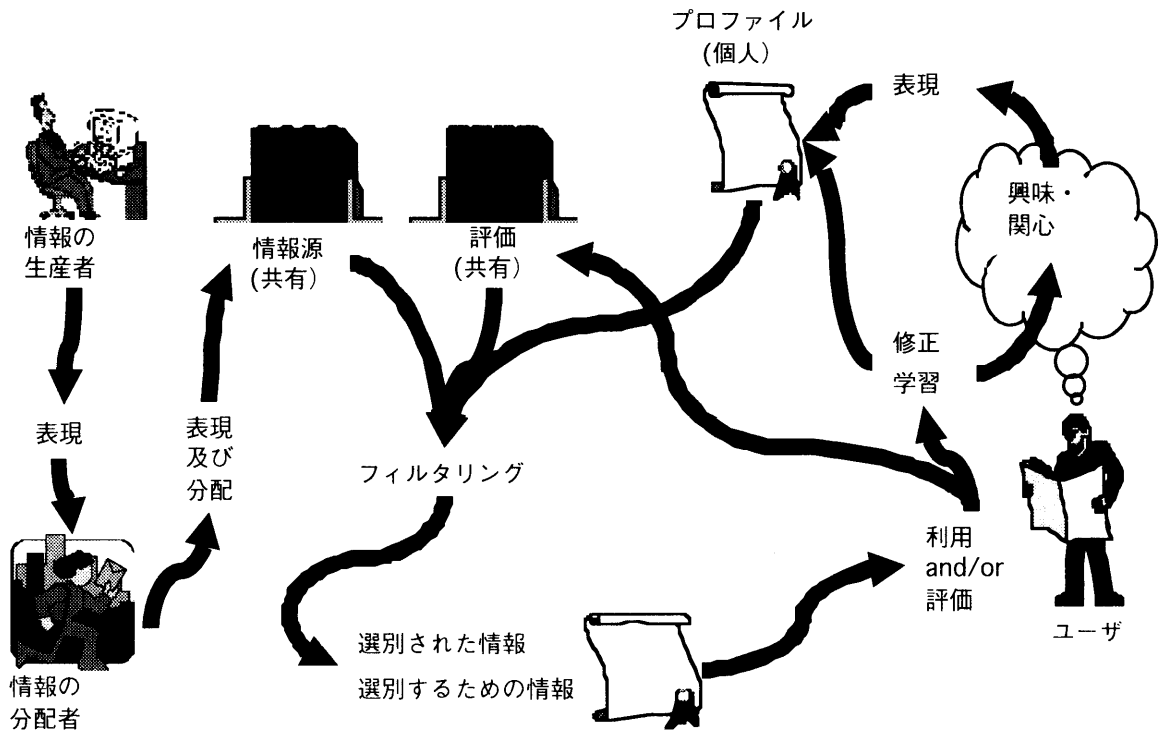

図1　情報フィルタリングの一般的モデル
Fig.1 General Model of Information Filtering

図1に与えた情報フィルタリングの一般的モデルについて説明する．この情報フィルタリングの考えは，インターネット上の情報に対してのみ有効なのでなく，多量の情報が存在する応用に関して適用可能な汎用性の高い考え方ではあるが，以下の説明ではインターネット上の情報を例として取り上げて説明を行う．

（1）情報の生産者：一般の利用者，企業，新聞社，など多岐多様に渡る．情報の生産者は，自分の持つ情報を何らかの形で表現して，情報の分配者に渡す．

（2）情報の分配者：情報の生産者から渡された情報をまとめて，情報源としてインターネット

上に分配する．元の情報を加工して表現しなおすこともある．WWWサーバーの提供者がこれにあたる．情報の生産者が情報の分配者のこともある．

（3）情報源（共有）：インターネット上でインターネットの利用者に共有される情報のあつまりである．

（4）評価（共有）：インターネット上の情報源の中の情報に対する種々の観点からの評価をあつまたもので，情報フィルタリングを行うときの材料として，利用者の間で共有する．

（5）プロファイル（個人）：ユーザ（利用者）がどのようなことに関心を持っているか，興味を持っているか何らかの形で表現したものをいう．情報フィルタリングを行うときの材料とする．利用者に適した情報フィルタリングを行うために個人毎に用意する．情報フィルタリングによって選別した情報がユーザ（利用者）が利用したり評価したりした結果によって，プロファイルは修正されたり学習を行ったりする．

（6）フィルタリング：情報源の中の情報に対して，利用者間で共有されているその情報に対する評価，および／または，各利用者のプロファイルを用いて選別を行う．フィルタリングの結果として，評価の高い情報，および／または，利用者の関心・興味にあった情報が選び出される．情報に対する評価と各利用者のプロファイルは，両方を同時に利用する方法も考えられるし，どちらか片方のみを利用する方法も考えられる．システム，すなわち，機械が，勝手に情報の良し悪しを判断するのが望ましくない場合には，利用者自身が情報を選別するための材料となる情報を提供することも重要である．

このような情報フィルタリングのモデルにより，『膨大な規模の』情報源の中から『有用な』情報を『その情報が有用であることを自動的に推測して』『自動的に』選別することが可能になると考えられる．

2．2　フィルタリングの方針

以下では情報フィルタリングを応用したシステムの研究または開発を行うときに検討すべき事項について考察を行っていく．最初は，情報フィルタリングの技術を何をフィルタリングするために用いるのかという方針についてである．大きくは次の二つに分けられる．

（1）有用な情報を選別して取り出す．

（2）有害な情報を選別して除去する．

前節までの説明では，有用な情報を取り出すことに限って話をすすめてきたが，情報フィルタリングは有害な情報を除去するたのにも用いることができる．有用な情報を取り出す方針を採用した情報フィルタリングの応用したシステムが，情報推薦システムである．一方，有害な情報を除去する方針を採用したものとして，Microsoft社のWWWブラウザInternet Explorer で使われているRACSiがある．

さらに，どの程度まで選別する情報を絞り込むのかによって，システムの実装は大きく異なることが予想される．

有用な情報を選別して取り出す場合では，方針は次の二つに分けられる．

（1－1）有用な情報のみを取り出す．：不用な情報は決して取り込まない．有用な情報が多少取りこぼしてもしかたがない．

（1－2）有用な情報を取りこぼさない．：不用な情報を多少取り込むことはしかたがない．

有害な情報を選別して除去する場合では，同様に方針は次の二つに分けられる．

（2－1）有害な情報のみを除去する．：無害な情報は決して排除しない．有害な情報を多少取り込んでもしかたがない．

（2－2）有害な情報は必ず除去する．：無害な情報が多少排除されてもしかたがない．

フィルタリングの方針によって，システムの実装は大きく異なる．何を優先して考えるかが重要な判断基準となる．

2．3　評価の種類

情報源の中の個々の情報に対する評価について考察を行うと，情報に対する評価は次の2つに分けられる．

（1）直接的な評価

（2）間接的な評価

ここで，直接的な評価とは文字通りユーザが直接入力したその情報に対する評価のことである．直接的な評価を利用するためには，利用者が情報を見るたびに利用者がその情報をどのように評価したかを利用者に何らかの方法で入力してもらう必要がある．利用者にとってその手間に見合うだけの直接的な利益を得られなければ，そのシステムの利用を止めるか，または，見た情報に対して評価を入力することを止めてしまうことが容易に予測される．

一方，間接的な評価とは，利用者または利用者グループの情報に対する振る舞いなどから推測して得られる評価のことでり，その情報へのアクセス回数、読んでいる時間、などのことである．間接的な評価を利用したシステムのよいところは，利用者にとって入力の手間がいらないということである．間接的な評価は，その指標が本当にその情報に対する評価になっているかどうかということを検証する必要がある．

利用者の手間のことを考えると，情報に対する評価として妥当な間接的な評価をみつけることが重要である．しかしながら，間接的な評価だけにするのではなく，利用者が他の利用者または自分のために情報に対して評価をつけたいと考えたときにはいつでも直接的な評価をつけられるようになっていることが望ましい．

2．4　情報フィルタリングの分類

最後に，アプローチの違いによる情報フィルタリングの分類を示す．

（1）内容に基づくフィルタリング

最初は，内容に基づくフィルタリング（context-based filtering ）であり，認知的フィルタリング （cognitive filtering ）と呼ばれることもある．これは，利用者のプロファイルと情報の内容がどれだけあっているかということに基づいて情報フィルタリングを行う方法である．情報の内容を解析するために，自然言語処理，および／または，自然言語理解の技術が応用されることがある．MIT のMaesらのNEWT[4] ほか多数の研究がある．

（2）経済的フィルタリング

次は，経済的フィルタリング （economic filtering ）であり，情報が持つ利益とその情報を得るためのコストとの相互関係により情報を選別する方法である．インターネットの現状ではWWW ページの情報はほとんどが無料なためか経済的フィルタリングに関する研究事例は見られない．

（3）協調的フィルタリング

最後は，協調的フィルタリング （collabotive filtering ）であり，社会的フィルタリング（social filtering ）とも呼ばれることがある．他のユーザの情報に対する評価を使って情報をフィルタリングする方法である．MIT のMaloneらのGroupLens [5]，著者の研究室の「やじうまくん」[6] ほか多数の研究がある．

4．まとめ

インターネットの現状を数値として示し，インターネット上には「品質の保証のない，膨大な情報が無秩序に存在する」という情報洪水の問題について述べた．この問題を解決あるいは軽減するたの技術である情報フィルタリングについて一般的モデルを示し，情報フィルタリングの研究または開発において検討が必要な事項の検討を行った．

良い悪いとは関係なく，良いところ悪いところも抱えたまま，インターネットはこれからますます発展し規模を拡大していくものと予想される．健全な社会の一部としての健全なインターネットが構築されるように，情報フィルタリングの技術を含めたインターネット技術の研究に取り込むことが重要であると考えている．

参考文献

[1] 日本インターネット協会編，インターネット白書'97，インプレス
[2] 日本インターネット協会編，インターネット白書'98，インプレス
[3] Alta VistaのWWWページ：http://www.altavista.digital.com/
[4] Pattie Mase：Agentthat ReduceWork andInformation Overload, CACM, Vol.37, No.7, pp.30-40 , 1994, URL:http://pattie.www.media.mit.edu/people/pattie/CACM-94.p1.html
[5] Paul Rensnick , Neophytos Iacovou Mitesh Suchak, Peter Bergstrom , John Riedl : GroupLens：An Open Architecture for Collaborative Filtering ofNetNews, Proc.of CSCW'94, pp. 175-186. , 1994.
[6] 杉井俊彦，北英彦，林照峯：アクセス回数を利用したWWWの人気ホームページ道案内システム，情報処理学会研究報告，97-GW-21 , pp.235-240 , 1997.