情報洪水の緩和のための
インフォメーションスケーリングの実現

慶應義塾大学環境情報学部徳田研究室
川又 浩一
kawauso@sfc.wide.ad.jp
指導教官
徳田 英幸
村井 純

平成8年 12月 21日

Abstract

コンピュータネットワークを通じて流入して来る情報の量は、人が処理できる許容量を超えてしまった。現在のところ、この情報洪水という問題に対しては、流入路の制限という解決策しかない。

ネットワークエージェントによる情報の自動受発信や、インフォメーションフィルタリングによる情報の自動選択などの実現は、この問題に対する重要な打開策となる可能性を持っている。

自動的に情報の受発信を行うネットワークエージェントは、人間の情報の受発信という行動を一部代行する。高度な知的処理を必要とするため、多くの技術課題が残っており、実用目的に使用するにはまだ研究が必要である。

インフォメーションフィルタリングにおける技術課題は、データベース分野における技術課題と類似点が多い。多くの研究の成果がそのまま利用できるため、テキストベースのいくつかのインフォメーションフィルタリングシステムが、ネットワーク上ですでに実用化しつつある。

代表的なインフォメーションフィルタリングに、プロファイル方式による、情報の自動選別がある。プロファイルとして好みのキーワードや、好みの文書を指定すると、コンピュータがこのプロファイルと流入して来る文書を比較し、重要度を計算する。コンピュータは、この比較の結果の評価値を見て、あるしきい値を越えない文書は破棄し、越えたものを人へ手渡す。

この比較のプロセスが未熟であると、必要な情報が捨てられたり、逆に必要でない情報が多く残ってしまったりする。

現在インフォメーションフィルタリングの研究者は、主にプロファイルの生成技術と、これと流入して来る文書を比較する技術に焦点を置いて、精度や速度の向上をはかっている。

本研究では、比較をして評価値を得たあとの、コンピュータが人に情報を渡す際のプロセスに着目した。そして本論文では、比較のプロセス精度の問題を解決する新しい情報ろ過の方法、インフォメーションスケーリングを提案した。

Abstract

Presently the amount of information that flows through the wide-area information system, has far exceeded the human capacity to transact.

At the moment, the only resolution to this information overflow problem is to restrict the amount of information lines.

Another would be, the realization of Network Agents that can make selective information retrieve possible. This can be an answer to the problems related to the overflow.

Because the act of automatic info-retrievment/dispatchment substitutes for human intellectual act, much study and research is necessary to overcome the technical difficulties, as well as bringing it to practical use.

The technical problems dealt with information filtering is similar in many ways with that of the database field. Because most of the research, especially in the information filtering of text-based documents can be used directly, there are already some services on the network that make use of this.

One typical information filtering would be the profile system where the information is automatically selected and sorted out. The computer then compares this profile to each incoming information, and calculates the importance and comes up with a score.

If the score is over the threshold, it is passed on to the client. If not, it is dismissed. If this filtering process is unstable and immature, there is a possibility that important information will be thrown away or cause an overflow of unnecessary information.

Presently, research efforts on information filtering is focused on the making and effectiveness of filtering. They are attempting to provide more accurate and fast information filtering.

In this research, We focused on the process in providing requested information to the client from the given profile. In this thesis, we proposed a new filtering scheme "Information Scaling" which improve the accuracy and efficiency of the filtering.

目 次

図 目 次

第 1 章

序論

1.1 はじめに

コンピュータと、それをつなぐネットワークの普及が進み、個人が簡単に、世界へ向けて情報を発信できるようになった。それらの情報の量は、ネットワークが社会へ浸透するとともに、増加を続けている。

コンピュータを通じて入手できる情報の量は、今では人が処理できる量と比べ桁違いに多くなった。これが情報洪水という現象である。

この現象に直面した人々は、増え続ける情報に対する嫌悪感を抱きながら、一方でそれらの中にあるはずの、自分にとって価値の高い情報に対する要望も持っている。

ある情報の入手経路において、価値の高い情報を選び出すためのコストが、情報を入手したことによって得られる報酬と比べて増大しすぎると、人はその経路を使った情報の入手をやめてしまう。

このコストを下げるための技術があれば、情報洪水を緩和できる。

1.2 本研究の目的と方法

本研究の目的は、人が量の多い情報を容易に扱う方法を考案し、情報洪水を緩和することである。このために、大量の情報のなかから、価値の高い情報を、人が見つけ出すことを支援する環境を設計する。

本論文では特に、コンピュータを使った情報取得について論じる。大きく分けて以下の 3 つについて述べる部分からなっている。

1. 現在一般的な情報取得
2. インフォメーションフィルタリングを使った情報取得
3. インフォメーションスケーリングを使った情報取得

まず、現在一般的なツールを使った情報取得の実例を挙げ、それらを比較検討する。次に大量の情報を適切な量に減らす、インフォメーションフィルタリングと呼ぶ技術について、最近の研究状況について述べる。次にインフォメーションフィルタリングで採用している情報ろ過のモデルの問題点を指摘し、変更を加える。変更後のモデルを、インフォメーションスケーリングと呼ぶことにし、その応用アプリケーションを提示する。

第 2 章

コンピュータによる情報取得

2.1 メール、ネットニュースを使った情報取得

ネットワークからの情報取得のためのツールとして、現在多くの人が使っているのがメールと、ネットニュースである。

2.1.1 MHE

インターネット上の情報提供サービスの一例として、メーリングリストがある。何百というメーリングリストがあり、様々な分野をカバーしている。自分の興味のある分野があれば、そのメーリングリストに参加することで、メールを通じてその分野に関する情報を得ることができる。

メールを読むための代表的なアプリケーションとして、エディタ emacs のフロントエンドである MHE がある。メールを到着順番に、ひとつずつ表示することができる。指定したキーを押すと、次のメールを表示するようになっている。次々とキーを押すことで、メールを読んでいく。

2.1.2 GNUS

インターネット上の電子掲示板の役目を果たしているのがネットニュースである。情報取得のためのツールの性質としては、メーリングリストに似ている。ネットニュースを読むための代表的なアプリケーションとして、エディタ emacs のフロントエンドである GNUS がある。

GNUS では、トピック別に、投稿された情報を到着順番に、ひとつずつ表示することができる。指定したキーを押すと、次の情報を表示するようになっている。次々とキーを押すことで、情報を読みすすむ。

2.1.3 MHE、GNUS の特徴

MHE、GNUS に代表するメールやニュースの表示アプリケーションは、図 (2.1) に示すような構造で表示を行っていることが多い。

情報のタイトルが上部に並んでいる。そのうちのひとつが選択できるようになっており、その情報の本文は下部に表示する。

メーリングリンスト、ニュースともに情報の取捨選択には、図 (2.2) に示すようなトピックの分類を使うことができる。

このトピックで情報を特定して、ニュースやメーリングリストを読んでいても、本当に興味を持って読める情報に会える頻度は、それほど高くない。

2.2 ハイパーテキストアプリケーションによる情報取得

MHE や GNUS は、情報を到着順に並べて表示する。これに対し、これから述べるようなハイパーテキストアプリケーションでは、情報を意味の関連によって並べ、図

```
   4  12/12 Yutaka Imai        pattern jouhou ron <<パターン情報論履修者の皆様
■  5+ 12/16 Yutaka Imai        pattern jouhou ron report <<パターン情報論TAの今
   6  01/17 Yutaka Imai        pattern jouhou ron <<パターン情報論TAの今井豊で
 -かんな[--]J_JJ:--%%-{+pattern}  6 msgs (1-6)      (mh-e show)----Bot----------
             t91279sc@sfc.keio.ac.jp,
             t91280yc@sfc.keio.ac.jp,
             t91318mn@sfc.keio.ac.jp,
             t91322hn@sfc.keio.ac.jp
Cc:          ishizaki@sfc.keio.ac.jp,
             t91488sy@sfc.keio.ac.jp,
             yim@sfc.keio.ac.jp
From:        Yutaka Imai <yim@sfc.keio.ac.jp>
Subject:     pattern jouhou ron report
X-UIDL:      787564202.001

パターン情報論TAの今井です。

パターン情報論の学期末のレポート課題についてお知らせします。
```

Figure 2.1: MHE, GNUS 等が採用している表示方法

```
       195: sfc.community.www
        23: sfc.courses.ipl2n
         7: sfc.official.cdp
         3: sfc.official.cns
         6: sfc.official.media-center
         9: sfc.official.shonan-fujisawa-gakkai
        45: sfc.official.wellness
         1: sfc.official
         1: sfc.academic.comp.c
      1146: fj.os.linux
        15: sfc.academic.comp.lisp
         1: sfc.academic.comp.pascal
        15: sfc.academic.comp.sas
         1: sfc.academic.comp.spreadsheet
         7: sfc.academic.ecology
         2: sfc.academic.france
         3: sfc.academic.interface
         5: sfc.academic.science
        64: sfc.community.cns
        37: sfc.community.co-op
         9: sfc.community.complaints
        10: sfc.community.jimushitu
         1: sfc.community.market.laptop
```

Figure 2.2: ニュースのトピック一覧

(2.3)のように表示する。

ユーザがその連結を選択してゆくことによって、次々と情報を読みすすむ。このような構造をハイパーテキストと呼び、表示するツールをブラウザと呼ぶ。

2.2.1 パッケージソフト

ハイパーテキスト構造は、とくに CD-ROM などを用いて大量のデータを配布する際によく用いる。各種のブラウザがあるが、基本的には図 (2.3)のような、画面に並んだボタンを押して次の画面に移る、という構造を持っている。

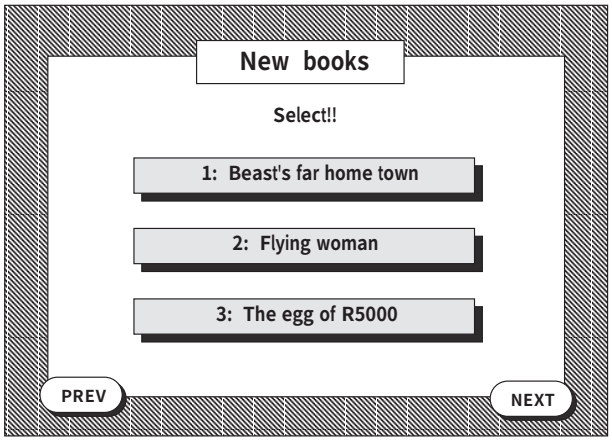

Figure 2.3: ハイパーテキストアプリケーションの例

2.2.2 WWW

ネットワーク上に蓄積してゆく情報を、ハイパーテキストの手法を応用して配布するシステムが WWW である。マウスによる連結先の選択だけで、次々とネットワーク上にある情報が引き出せる。

2.2.3 ハイパーテキストを使ったアプリケーションの特徴

情報をハイパーテキスト構造にしておくと、情報の間の飛び移りが容易にできる。マウスによる連結先の選択だけで、次々と関連する情報が引き出せる。

しかし、情報を読み進む方向が、例えば MHE では前と後という二つの方向しかなかったのに対して、ハイパーテキストのアプリケーションでは、無限の方向があり、しかもそれらが相互に絡み合っている。

したがって例えば Mosaic では、情報を読み進むのに頻繁なマウス操作を必要とする。

2.3 問題点

一度に表示する情報がひとつである。この点では、MHE も Mosaic も変わらない。

メーリングリストを積極的に利用しようと思えば、一日のメールの量が数百通を越える。これらのメールひとつひとつ全てに目を通すには、多くの手間と時間が必要である。MHE のようなアプリケーションは、ひとつひとつのメールを逐次表示する構造になっているため、少なくとも一回は全てのメールを表示しないと、全体に目が通らない。

Mosaic でも、少なくとも一回は全てのページを表示しないと、全体には目が通らない。さらに、どの情報を読み終ったら情報を全て読んだとするのか、判断するのが

難しい。ハイパーテキスト構造の情報は、意味による情報間の移動を実現しているかわりに、自分が今見ている情報が、全体の中でどの位置にあるのか、分かりにくくなるのである。

MHE でメーリングリストのメールを読んだり、Mosaic で WWW のページを読んでいる様子は、砂場に落ちたピンをルーペで探している様子に似ている。

そして、コンピュータネットワークの拡大、その上で情報を発信するユーザの増加によって、探すべき砂場は広がるばかりである。しかしルーペの視野は変わらない。これが情報洪水と呼ぶ状況である。

第 3 章

インフォメーションフィルタリング

3.1 インフォメーションフィルタリングによる情報取得

メーリングリストやネットニュースへの、積極的な参加を控えているといういう人は多い。情報洪水を解決するには、現在のところ、情報の流入路の取捨選択をするしかないからである。

数値計算が主な仕事ではないコンピュータは増えている。それらのコンピュータは、情報の入力、編集、配布などに使う。

しかし、情報が伝えたいこと、情報の内容の処理については、コンピュータは感知していない。そこで現在、コンピュータが情報の内容を判断して、知的に動作するようにする研究が始まっている。ネットワークエージェント、そしてインフォメーションフィルタリング技術の研究である。

ネットワークエージェントとは、情報の受発信という人の行為の一部を、コンピュータが肩代りするシステムである。ネットワークエージェントは、ネットワーク上を渡り歩いて情報を探索する。主人が必要としている注文などを代わりにとる。

そのような行動をするシステムは、一種の人工知能である。人工知能に関する現在の技術水準では、人と同じレベルの知的活動をするようなネットワークエージェントを実用化するまでには至っていない。ネットワークエージェントの研究には、まだ多くの課題が残っている。個人が、情報洪水への対策としてネットワークエージェントを利用するようになるには、まだ時間がかかるであろう。

もうひとつの解決策である、インフォメーションフィルタリング技術とは、流入して来る情報の中から、重要と思うものをコンピュータが推定し、それらを優先して人に渡すための技術である。

情報の入手にかかる時間における、情報の選別が占める割合を少なくすることが、インフォメーションフィルタリングの主な目的である。

インフォメーションフィルタリングを利用するユーザは、情報に対する自分の興味を、システムが定める形式に従って記述した、プロファイルと呼ぶ要求にまとめてフィルタリングシステムに提出する。

コンピュータは、その個々のユーザが提出したプロファイルと、流入して来る情報を見比べ、ユーザにとって必要と判断した情報だけを手渡す。

3. 2 システム構成

インフォメーションフィルタリングを実現するには、コンピュータは新しい情報を、流入して来る速度よりも早く選別し終えなくてはならない。

情報を知的に扱って内容を理解し、それに基づいてフィルタリングをしていては、現在の技術水準では実用的なシステムができない。そこで多くのシステムは、テキストの表層的な構造を利用してフィルタリングしている。具体的には、テキストの全文検索や、部分列検索などの、データベース分野で培って来た技術を利用する。

図 (3.1) に示すのは、代表的なインフォメーションフィルタリングのモデルである。

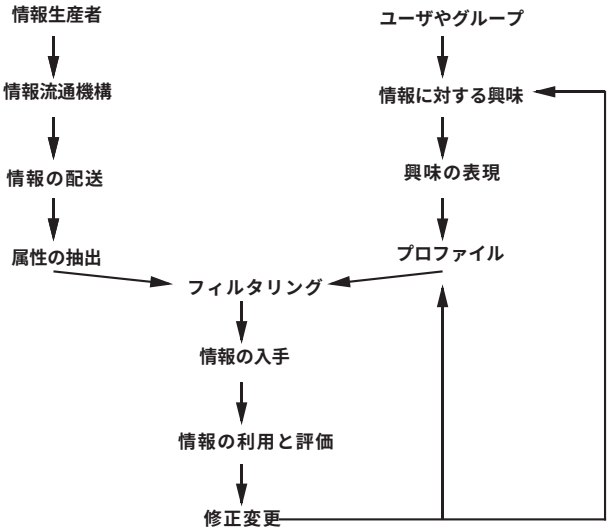

Figure 3.1: インフォメーションフィルタリングのモデル [Mor93]

インフォメーションフィルタリングの研究者は、とくに検索の過程と、ユーザの要求プロファイルの生成技術に着目し、その精度や速度の向上のための研究を進めている。これらの研究内容について、以下に概要を述べる。

3.2.1 テキスト検索技術

多くのインフォメーションフィルタリングの研究者は、プロファイルと流入して来る情報をマッチングする過程に焦点を置いて研究している。この過程には、データベース分野で培って来たテキストの検索技術を応用する。

一般的なデータベースにおけるテキスト検索

データベースの研究者のテキスト検索に関する研究成果に、キーワードの列と論理演算子を用いた Boolean 検索法や、Vector Space Model を使った検索法などがある。

これらの技術は、そのままインフォメーションフィルタリングにも応用できる。

Vector Space Model は、文書の類似度を計算するためのモデルである。文書を全て単語に分解し、各単語の出現回数を数える。そして文書をその単語の数だけの要素からなるベクトルで表す。二つの文書の類似度を、そのベクトルの一致の度合として計算する。一致の度合は数値として算出できるので、定量的に文書の特徴を扱うことができる。

この技術は、とくに自然言語のプロファイルを利用したフィルタリングに応用できる。

一度キーワードの指定によるフィルタリングをして得られた文書を使って、次々にそれに似た文書を探すという、フィードバック機構の実現にも用いることもできる。こ

れを relevance feedback といい、多くのフィルタリングシステムが採用している機構である。

日本語テキスト検索

英文は、図 (3.2) に示すように単語ひとつひとつを空白で区切っている。このために単語の抽出が容易である。日本語の場合は単語の区切りが明白でない。したがって英文よりも単語の抽出は困難である。

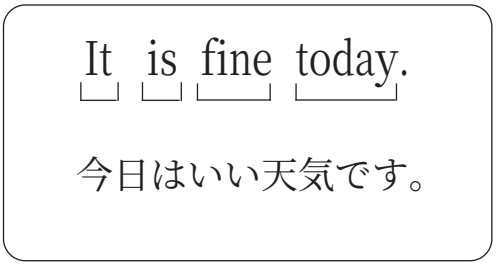

Figure 3.2: 英語と日本語

この日本語の単語抽出の問題は、自然言語処理分野での研究によって、ほぼ解決している。研究成果のひとつとして、形態素解析システム JUMAN がある。単語抽出が可能なフリーソフトウェアとして広まっており、日本語のインフォメーションフィルタリングの研究者が積極的にこれを利用している。後で述べる WAIS でも、このシステムを利用して、日本語を Vector Space Model で検索可能にしている。

プロファイルと流入して来る情報の比較に重点をおいて構成した、代表的なインフォメーションフィルタリングシステムをいくつか挙げる。

Information Lens

半構造化したテキストのフィルタリングを行って、グループによる情報の共有などをサポートするシステムである。あらかじめ構造を持たせることができる情報については、かなり有効なフィルタリングが出来ることを示した。

WAIS

情報を提供する WAIS サーバと、多数のプラットフォームで利用できる WAIS クライアントからなる分散情報提供システムである。日本語による情報を含めた、数百のサーバが立ち上がっており、それぞれ異なるテーマの情報を提供している。

WAIS は、高速な検索を行うために、あらかじめサーバにあるテキストからキーワードを抽出して、インデックスを作成しておくという方法を使っている。ネットニュースの記事、メール、TeX のソースファイルなどの、もとになる情報の種類によってキー

ワード抽出の方法を変えてインデックスを作成している。動画や音声などのマルティメディアデータに関しては、ファイル名をインデックスに利用している。

クライアントから検索要求を送ると、抽出しておいたインデックスを使って検索を行う。サーバは、クライアントが送って来たキーワードの出現回数をカウントし、多い情報に高いスコアを与える。そしてその情報をもとにフィルタリングする。または、クライアントから送って来た文書から抽出した、キーワードの分布の比較し、同じ単語が多く出現しているものに高いスコアを与える、vector space model を用いてフィルタリングすることもできる。

SIFT

WAIS と同様なテキスト検索技術を用いて、ネットニュース等に流れる情報をフィルタリングするシステムである。その検索の高速化に焦点を当てて実装をしている。ひとつのサーバで 7500 人分以上のプロファイルを管理、そのそれぞれについてネットニュースからフィルタリングを行い、一日一回のペースで結果を図 (3.3) に示すようなメールに梱包して配送することができている。

```
=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=-=
Subscription 1: skimming agent no...

Article: comp.lang.c++.71562
Message-ID: <cbb_9502010144@trisoft.com>
From: Don.Hankins@f1.n328.z1.fidonet.org (Don Hankins)
Subject: Re: Borland C++ V4.02 IDE Debugging
Score: 100
First 40 lines:
 From: deh@fir206.cray.com (Don Hankins)
 Message-Id: <1995Jan27.082329.22396@driftwood.cray.com>
 Date: 27 Jan 95 08:23:28 CST
 The Borland Debugger manual that comes with the compiler describes how
 to debug DOS targeted programs.  I was just skimming through it last night.
 You don't use the IDE interface for debugging DOS targeted programs.  Instead
,
 you run the debugger under DOS, using a command line.
 -Don Hankins
 deh@cray.com
 Disclaimer:  Opinions are my own, and not those of my wife, kids, dog, cat,
 or Cray Research.
 In article <3g7fmo$l16@eplet.mira.net.au>, gibney@werple.mira.net.au (John
 Gibney) writes:
 > 'lo everyone...
 >
 > Just tried debugging a simple DOS-targetted program in the IDE...  but it
 > says I can't debug anything except Win. programs with the integrated
 > debugger in the IDE.  Surely this is pretty silly...  Is there something
```

Figure 3.3: SIFT によるネットニュースのフィルタリング結果

プロファイルの提出には、メールと、図 (3.4)に示すような WWW クライアントを使用する。

3.2.2 プロファイル生成技術

インフォメーションフィルタリングサービスを利用するユーザは、自分が何を求めているのかをシステムに教える必要がある。要求は多くの場合図 (3.4) のようなプロファイルの形をとる。

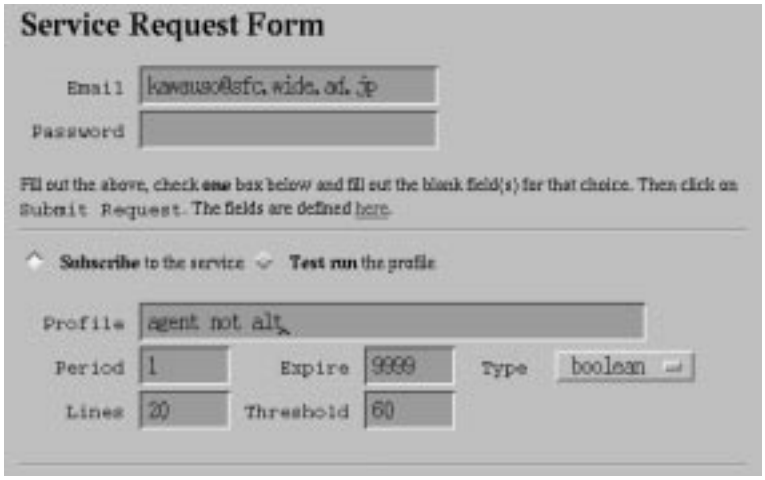

Figure 3.4: SIFT サーバへのプロファイル提出例

プロファイルが正確であれば、検索の結果も好ましいものになる。したがってこのプロファイルの作成過程は、インフォメーションフィルタリングの研究のもうひとつの焦点である。

この過程に着目して構成したインフォメーションフィルタリングシステムをいくつか挙げる。

chknews

chknews[Mor94] は、コンピュータがプロファイルを自動作成する、ネットニュースのフィルタリングシステムである。chknews は、ユーザがニュースを読む際に、どの記事に時間をかけたかを観察する。興味のあるニュースはより時間をかけて読むという仮定をする。注視時間の長かった記事をプロファイルとして記録して、これを参照しながらフィルタリングをする。

Autodesk

Autodesk[Bac92]は、人の興味を推測する過程に焦点を置いた、プロファイル自動生成システムである。競合エージェントとして、いくつかの機械学習アルゴリズムを組み合わせ、そのエージェントにユーザの興味を学習させてプロファイルを生成する。遺伝的アルゴリズムやニューラルネットを利用した学習アルゴリズムを実現している。

Lyric-Timesystem

Lyric-Timesystem[Loe92] は、人が普段さりげなく聞く、音楽のような情報を、フィルタリングするためのシステムである。このようなくだけた情報に関するプロファイルを厳密に作成するのは難しい。そこで普段音楽の再生をしながら、ひとつひとつの

音楽に関する感想を記録し、プロファイルを自動作成する。この際、例えばそのときのユーザの気分などの情報も聞いておく。楽しいとき、悲しいとき、落ち着きたいとき、などのそれぞれの気分に別なプロファイルを作成し、これら複数のプロファイルを弾力的に運用することで、さりげないフィルタリングを行うシステムを提供する。

3.3 インフォメーションフィルタリングの問題点

3.3.1 SIFT に見る問題点

SIFT を使った情報取得のいくつかの特徴のうち、以下の 3 つに着目した。

- あらかじめどのような情報が得られるのか予測できる
- 情報の取捨選択の判定の過程が見えない
- プロファイルを作るのが難しい

この 3 つの特徴のそれぞれについて、以下に問題点を挙げる。

情報の予測ができる

SIFT ではあらかじめユーザが指定したプロファイルに一致した記事を送ってくる。したがって、SIFT のユーザは、新しく着いたメールを読む前から、そのメールがどんな内容であるか予測できる。

手間をかけて MHE などのツールを使ってメールを読み進む理由のひとつとして、そこに何が書いてあるのか分からない、ということがあげられる。新しい情報への期待の半分位はこの不確定要素が絡んでいる。SIFT が送ってくるフィルタリングした情報には、この要素がない。

したがって、役に立つ情報が書いてあることが分かっているのにもかかわらず、他のメールが多かった場合、それらのメールと比較した、SIFT のメールの優先順位が下がる。

その結果として、SIFT のユーザは、せっかくのフィルタリング結果を読み飛ばしてしまう可能性がある。

フィルタリング過程が見えない

3 つの記事が SIFT のメールの中に入っていたとする。それを受け取ったユーザからは、SIFT がその記事を、一体どれくらいの量の記事のなかから選び出して来たのか、どうしてその記事を選んだのか、そのメールを読んだだけでは分かりにくい。

ここで、ユーザがフィルタリングシステムを利用しないで、もし手で全ての情報に目を通し、その後で振り返ってみればその 3 つの記事が最も面白かった、あるいは役に立ったという結果になったとする。この例は、フィルタリングシステムを利用して 3

つの記事を得た前述の例と、結果は同じである。しかし、その 3 つに絞り込んだ過程がどのようなものであったのか、SIFT のユーザには見えていない。

その 3 つの記事にどれくらいの価値があるのか、手を使って苦労して選んだ場合には、その苦労度からなんとなく理解できる。しかし、フィルタリングシステムが自動的に、システムのユーザの見えないところで選んできた情報の場合は、その価値に実感がともなわない。

フィルタリングシステムが自動的に情報を選び出すと、その情報が、その他の情報を含めた全体像の中でどう位置しているのか、どれくらいの相対量を持っているのかが分かりにくい。

これだと、ユーザは、フィルタリングシステムが選んだ情報の価値を正しく判断できない。

プロファイルの作成が難しい

プロファイルについては、以下に挙げる問題がある。

- プロファイルを記述するのに手間がかかる
- 知らない事象については、プロファイルが記述できない

SIFT を利用するユーザは、プロファイルとして自分の興味を、単語の論理式などの形で記述しなくてはならない。この手間を惜しむと、興味のモデルが忠実でなくなり、フィルタリングの精度が甘くなる。

またシステムの矛盾点として、SIFT に送ってもらいたい情報を、先にプロファイルで記述するという順序の逆転がある。ある事象 A について、プロファイルを記述するには、事象 A についての予備知識が必要である。

フィルタリングシステムに送ってもらいたい情報を、プロファイルで正確に注文できる人とは、すでにその情報を知っている人ということになる。

3.4 問題の一般化

SIFT は、インフォメーションフィルタリングモデルの実装例として問題のないシステムである。これまで挙げてきた問題は、SIFT 固有の問題ではなく、インフォメーションフィルタリングのモデルが持つ問題であると考えられる。そこでここでは問題を一般的に整理しなおし、モデルの改善を図る。

MHE や GNUS における情報の流れは、図 (3.5) に示すようになっている。

I_1, I_2, I_3という情報があったとする。MHE　や GNUS　型の情報経路のユーザ M 氏がいたとすると、M 氏は I_1, I_2, I_3と線的に読み進んで全ての情報に目が通る。情報の表示順は、MHE　などは到着順であり、ハイパーテキストでは、意味によって決まる。

インフォメーションフィルタリングにおける情報の流れは、図 (3.6)に示すようになっている。

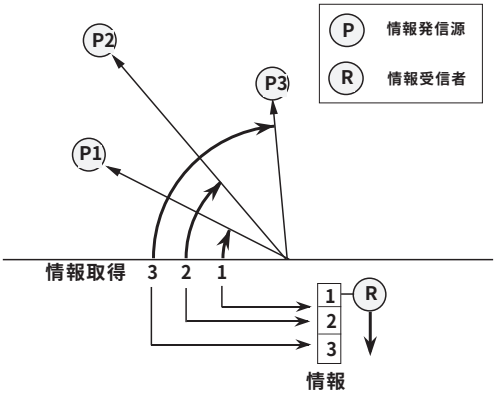

Figure 3.5: MHE や GNUS における情報の流れ

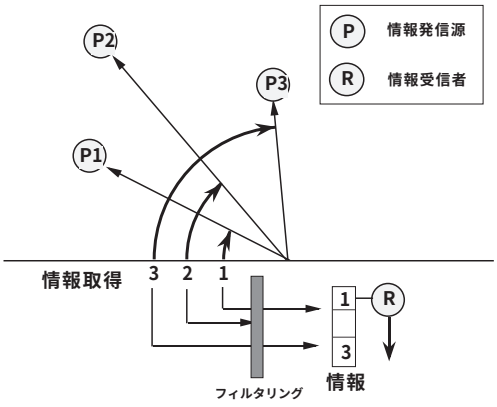

Figure 3.6:インフォメーションフィルタリングにおける情報の流れ

情報は、フィルタリングシステムを通過し、有用な情報だけになって、システムのユーザの手に渡る。

フィルタリングシステムのユーザ F 氏がいたとして、F 氏の渡したプロファイルから計算した、各情報のスコアは、$S(I_1) = 28, S(I_2) = 3, S(I_3) = 75$ であったとする。

フィルタリングシステムは、例えば有用な情報かどうかのしきい値が 50 であったとすると、スコアが 50 よりも高かった I_3 だけを F 氏に手渡す。

しかし、F 氏は、こうして手渡された情報 I_3の情報価値を正確に判断できない。

フィルタリングする前の情報を振りかえると、それらの情報全体には、

> **I_1, I_2, I_3, という情報があり、そのうち I_1, I_2は、これまで F 氏が関心を持ってこなかった種類の情報、そして I_3は今関心を持っている情報である**

という内容を含んでいる。フィルタリング後の情報は、これらの内容を含んでいない。したがって、F 氏には伝わっていない。

インフォメーションフィルタリングのモデルでは、コンピュータがもし完全なアルゴリズムで情報の取捨選択をしても、F 氏にとって満足な結果が得られないことになる。

現在研究が進んでいるテキスト検索技術、プロファイル生成技術が進んでも、この問題は解決しない。

3.5 解決方法

情報をフィルタリングする理由のひとつに、コンピュータと人間の間の情報経路の限界がある。ディスプレイに表示できる情報量や、同時に目に入れられる情報量は決まっている。

インフォメーションフィルタリングのモデルは、このボトルネックに合わせて作成しているともいえる。したがって、情報過多を解決するために、情報の表示過程に着目することは有効なはずである。

このことをさらに具体的に検討するため、以下のシステムを参照した。

3.5.1 情報表示関連モデル

インフォメーションフィルタリングとは若干異なる分野で、設計の参考になると考えられるモデルがある。以下にそのモデルのうち二つのものについて概要を述べる。

魚眼レンズモデル

人は、自分の近傍を詳しく表現し、もっと遠い領域に関しては主要な目標 (ランドマーク) のみを示す。魚眼レンズモデル [Fur86] は、この観察に基づいて作ったモデルである。

例として図 (3.7)に示すような「ロサンゼルス人の世界の眺め」という地図がある。この地図は遠近が実際の地図と比べて歪んでいるが、ロサンゼルス人の世界の最も重要な特徴が現れている扱いやすい省略図である。

ハイパーテキスト構造の情報には、現在表示している情報が、全体のなかでどのような位置にあるのかが、分かりにくいという問題点がある。魚眼レンズモデルの提唱者は、この問題にモデルを適用して解決しようという試みを行っている。

ハイパーテキストにおける、特定のノードやリンクを、表示するべきかどうかを決定するために、DOI(Degree of Interest関心度) 関数を用いることによって、魚眼レンズモデルを形式化する。DOI は二つの要素に分かれており、ひとつがノード (x)の API (apriori interest先験的関心) を表し、もうひとつが現在のノード (o)との距離 D(o , x) を表す。

この試みでは、DOI 関数の詳細や、実際の表示方法についてはとくに言及していない。

しかし、DOI 関数の先験的関心という項目は、インフォメーションフィルタリングにおける文書のスコアに相当すると考えることもできる。

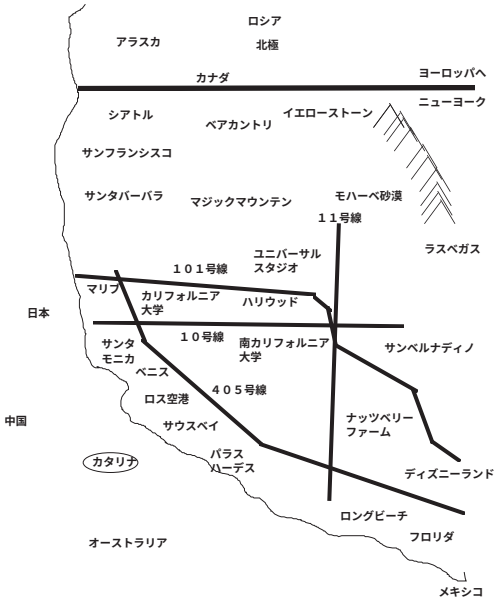

Figure 3.7: ロサンゼルス人の世界の眺め [Han91]

このようなブラウジングのモデルに、インフォメーションフィルタリングの技術を導入すると、モデルを実用化できる可能性が出てくる。

新聞

新聞は、長い歴史をもつ情報媒体である。情報の表示技術に関しては、すぐれた技術を経験的に確立している。新聞が採用している情報の表示モデルは、紙などに固定した情報の表示モデルとしては、最適なモデルのひとつである。[Ohk93]

コンピュータのディスプレイと比較して、新聞の紙面は数倍以上広い。紙面では、解像度の高い紙の特質を生かして、数個の情報を一度に表示することが可能である。

さらに、ひとつひとつの情報について、その内容の要約を分かりやすいかたちで並べている。情報は、ヘッドライン、そのバックグラウンドパターン、要約、本文に分かれている。バックグラウンドパターンを利用すると、「重厚」「まじめ」というイメージや、「疑惑的」「非日常的」など、情報の持つイメージを意図的に、人々に共通に伝えることができる。[Hat94]

さらに、重要度が高いと判断した情報は、紙面の中で全体的に大きく、そうでないものは小さく並べる。

このような工夫により、コンピュータ上の通常の情報の表示方法である、単純なテキスト表示と比べ、情報の受信者は、より早く必要な情報にたどり着ける。そして記憶に、より鮮明に残る。[Ohk93]

第 4 章

インフォメーションスケーリング

4.1 インフォメーションスケーリングモデル

魚眼レンズモデルでは、ハイパーテキストにおける現在値を示すために、ノード間の距離からスコアを求めて、結果によって情報の表示の詳しさを変えることを提案していた。これによりリンク構造の見晴らしを良くできる。

しかし全ての情報が、リンクによって結んであるわけではない。またインフォメーションフィルタリングのモデルによると、この距離は個人によって異なる。

そこでインフォメーションフィルタリングと魚眼レンズのモデルを組み合わせ、入手した情報の量と、現在表示できる情報の量を比較しつつ、個人の興味も参照しながら、情報表示の詳しさを適宜変更することを可能にするモデルを考案した。

この組み合わせモデルでは、情報のスコアを、情報の表示の詳しさを調節するために用いる。スコアを求める際には、ノード間の距離ではなく、情報の重要度ををを利用する。その際にはインフォメーションフィルタリングの技術を応用する。

この手法を本論文では**インフォメーションスケーリング**と呼ぶ。

インフォメーションスケーリングにおける情報の流れは、図 (4.1) に示すようになっている。

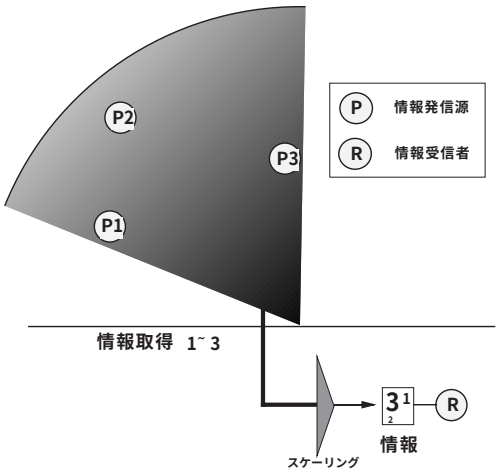

Figure 4. 1: インフォメーションスケーリングにおける情報の流れ

I_1, I_2, I_3 という情報は、スケーリングシステムを通過しても、その数は減らずに、システムのユーザの手に渡る。

コンピュータの情報表示能力 Dを 1 とすると、情報数が 3 であるので、同等に配分すれば $D_{I1} = 0.3, D_{I2} = 0.3, D_{I3} = 0.3$ とすることで全ての情報を表示できる。

ここで、スケーリングシステムのユーザ S 氏がいたとして、S 氏の渡したプロファイルから計算した、各情報のスコアは、$S(I_1) = 28, S(I_2) = 3, S(I_3) = 75$ であったとする。これを加味すると、$D_{I1} = 0.26, D_{I2} = 0.028, D_{I3} = 0.71$ とすれば全ての情報を表示できる。

4.2 インフォメーションスケーリングシステムの設計

インフォメーションスケーリングを使用した、テキスト情報の表示を行う試作システムの設計について述べる。

スケーリングとは、あるデータの流れから標本をとり、もとのデータの一部だけを表示する手法である。([Del93])

本システムで利用するインフォメーションスケーリングの利用モデルは、図 (4.2) に示すような構造を持っている。

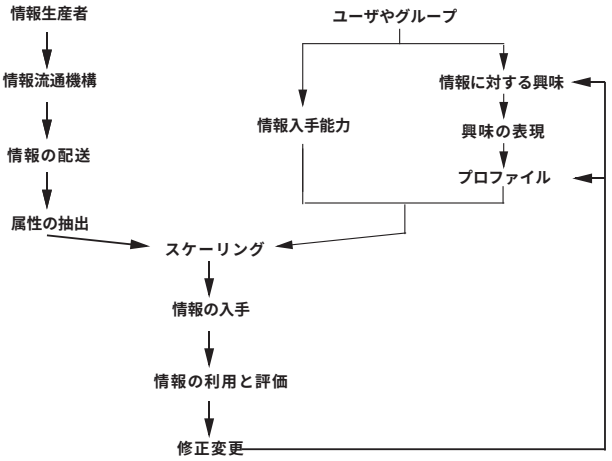

Figure 4.2: インフォメーションスケーリングのモデル

4.2.1 スケーリングする要素

本システムでは、スケーリングはテキストデータに適用する。各々の情報を、スコアの結果により、以下の要素を変更して表示する。

- 表示位置、順序
  スコアの高い情報を優先的に表示する。

- テキストレベル
  インフォメーションフィルタリングにおける、単語抽出の技術を利用して、文章のうち単語だけを抜きだして表示することで、決まった情報表示範囲に収まるようにする。文章は、画像と違って縮小しすぎると文字が見えなくなってしまい、情報価値がほとんどなくなってしまう。そこで、文章のうち単語だけを読めるサイズに保っておき、それ以外の部分をさらに縮小することで、見掛け上の情報価値を残したまま全体を縮小する。このレベルを本論文ではテキストレベルと呼ぶ。

4.2.2 スケーリングに用いる変数

3 つの要素をスケーリングするために、以下の 3 つの変数を用いる。

- 表示可能な情報の量
  指定した画面範囲の中で、どれくらいの情報が表示できるかを表す数である。
- 情報のスコア
  プロファイルを利用して求めた各情報のスコアである。

本システムではインフォメーションフィルタリングシステムと同様なプロファイルを作成し、それを利用している。

423 スケーリングの過程

ドキュメント

テキストであれば、どのようなファイルでもスコアを求めることができる。本研究では、ドキュメントのサンプルとして日本経済新聞朝刊の記事を約 800 個用意し、スコア計算に使用した。

新聞記事は、タイトル、日付、紙面での位置などの付加情報を持っている。これを構造化してドキュメントに格納し、スコア計算の際に用いた。

図 (4.3) にドキュメントの例を示す。

日米世界の中で（４）勝負アリ？マルチメディア――術力に勝る総合力。

九月十九日に京都で始まる国際電気通信連合（ＩＴＵ）の全権委員会議。閣僚級五十人をはじめ世界各国から千三百人を迎えるホスト国の日本は、地球規模の情報通信ネットワーク構築に向け、「政策会議」の設置を提案する。しかし、郵政省幹部の表情はさえない。肝心の米国が難色を示しているからだ。　「グローバル・インフォメーション・インフラストラクチャー」（ＧＩＩ）と呼ぶこのネットワーク構想を提唱したのは、もともと米国のゴア副大統領である。同副大統領は今年三月のＩＴＵ開発会議（ブエノスアイレス）でも「ＧＩＩを推進すれば、全世界で数千億ドル規模の経済成長の加速要因になる」と力説している。　その米国がＧＩＩを国際的に話し合う政策会議の設置に、なぜ難色を示すのか。ある郵政省幹部は「米国は自国の制度や政策を世界の標準として押し通すつもりなのだ。米国方式を全世界で採用すればいいという姿勢だから、各国との調整には当然、否定的になる」と指摘する。　米国は、情報・通信産業を二十一世紀を担う最重要産業と位置付けている。中でも家庭のテレビの上に装置を置き、ＣＡＴＶや衛星放送、ゲーム、映画、新聞記事など各種の情報を双方向でやり取りする「マルチメディア」は最有望株だ。　各国がある程度の技術を持ち様々な方式が併存するこの種の分野では、一つの方式のシェアが過半数を超えた時点で爆発的に成長する。かつてソニーとの競争を経て、日本ビクターなど松下電器産業グループが、ビデオのＶＨＳ規格を世界標準に育て上げた例をみでもわかる。

以下略

Figure 4. 3: ドキュメント例

スコアの計算

スコアの計算には、プロファイルに指定したキーワードの出現回数を数え、それらを総合してスコアを求める、全文検索法を使用した。プロファイルの書式は、図 (4. 4) に示すようになっている。

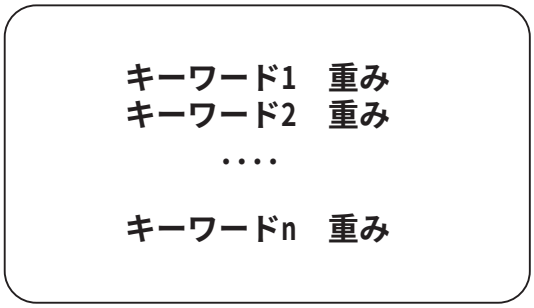

Figure 4.4: プロファイルの書式

キーワードは複数指定できる。具体的なプロファイル例としては、図 (4.5)なものが考えられる。

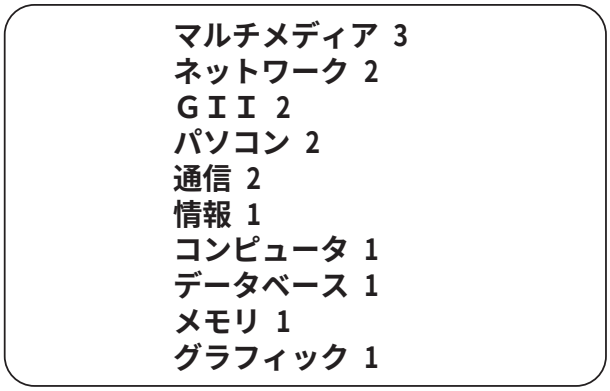

Figure 4.5: プロファイル例

そして、ひとつひとつのキーワードについて以下のような計算を行い、その結果を全て加算して、以下ドキュメントと呼ぶ、ひとつひとつの情報のスコアを求める。

$$\frac{log(N_{keyword}) * W_{keyword} * W_{document}}{S_{document}}$$

$N_{keyword}$はキーワードの出現回数、$W_{keyword}$はキーワードの重み、$W_{document}$ は、ドキュメントの作成段階でつけた、そのドキュメントの一般的な重要度、$S_{document}$ はドキュメントの長さである。

$W_{document}$ は、とくに指定がなければ 1 として計算する。

プロファイルのキーワードが、ドキュメント中にどれだけ登場しているか、カウントする。その結果に、重みやドキュメントの長さ等を加味し、スコアを計算している。

スケーリング表示

スコアの高いものを優先的に表示するため、ドキュメントをスコアの値の大きい順にソートする。例えば、図 (4.5) のプロファイルの場合、記事の優先度は図 (4.6)に示すような順になる。

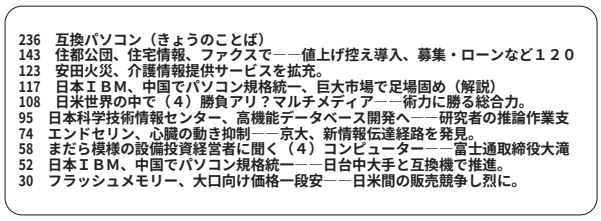

Figure 4.6: スコアの高かった記事

さらにテキストレベルの変更を行う。試作システムでは、ドキュメントから単語だけを抜きだして、それ以外の部分を削除することでレベルの変更をする。例えばテキストレベルの変更によって図 (4.7)は、図 (4.8)のようになる。

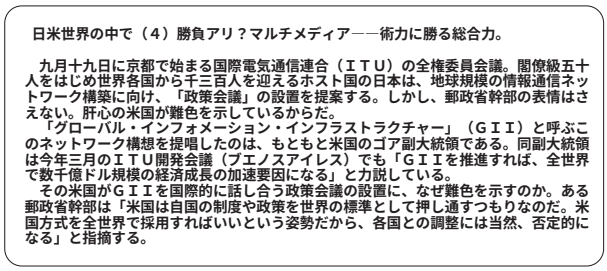

Figure 4.7: テキストレベル高

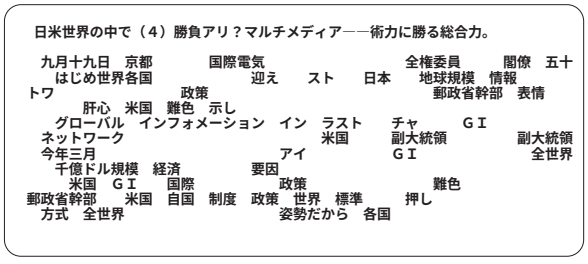

Figure 4.8: テキストレベル低

単語かどうかの判断には、単語辞書を利用している。試作システムの単語辞書は、主に名詞、固有名詞、数詞で構成しており、単語数は約 10 万語である。

第 5 章

結論

5.1 まとめ

これまでのインフォメーションフィルタリング研究動向を探り、どのような問題が解決に向かっており、どのような問題が残っているのかを考察した。

そして、インフォメーションフィルタリングのモデルが持っている欠点を挙げ、その改善モデルであるインフォメショーンスケーリングモデルを提示した。そして、それを応用した試作システムを作成した。

フィルタリングシステムは、そのユーザの目の届かないところで、フィルタリングの全過程を終了してしまう。そのため、ユーザは、フィルタリングの精度が十分であっても、手にした情報の価値を正確に判断できない。

フィルタリングの目的は、情報を入手する速さを、コンピュータと人との間の、情報が流れる速度の限界に合わせることである。

これに対し、スケーリングの目的は、コンピュータと人との間の情報が流れる速度の限界を、拡大することである。

スケーリングシステムでは、情報を人が全体像を把握しやすいように表示する。試作システムでは、ドキュメント表示の質を、その重要度に合わせて適宜変更して、スケーリングを行なった。スケーリング過程では、インフォメーションフィルタリングシステムにおける、単語抽出技術およびプロファイルとドキュメントの比較技術を利用した。

5.2 今後の課題

インフォメーションスケーリングのモデルを利用したシステムを、さらに改善するために、以下のような課題に取り組んでいる。

- スケーリングする要素の増加
  フォントサイズ、ウィンドウサイズ、表示する位置等の要素が考えられる。表示に用いる座標は通常 2 次元である。しかし次元数を増やせば、それだけスケーリングできる要素が増える。

- スケーリングサーバとクライアントの分離
  スケーリングの過程をサーバ内で行い、結果だけをクライアントに送信するようにすれば、情報転送の効率があがる。知的情報圧縮技術として、スケーリングシステムを利用できる。

- プロファイル作成環境
  本研究では、プロファイルの作成、更新については深く考察しなかった。しかし、インフォメーションスケーリングシステムにおける、プロファイルの占める役割は大きい。

  キーワードのシソーラスや、キーワードの関連辞書、概念辞書などを利用する方法が考えられる。

- 他システムとの提携
  現在人気のある情報、といった指定をできるようにするには、サーバと提携する必要がある。また、コンピュータにとって現在難しいインフォメーションフィルタリングも、人が行うことは可能である。したがって、新聞記者のような、フィルタリングをする人がいれば、それを前提としたフィルタリング + スケーリングシステムを作成できる。この二つのモデルを利用したシステムでは、好みの記者が選んだ情報を見たい、といった指定をすることで、フィルタリングを行い、さらに自分の好みでスケーリングをしながら読む、といったことが可能である。

謝辞

本研究を進めるにあたり御指導を頂きました、慶應義塾大学助教授の徳田英幸博士と村井純博士、そしてテキスト検索技術について御助言いただきました慶應義塾大学の石川直太博士に感謝します。

並びに同大学環境情報学部の徳田・村井研究室の諸氏、および同大学プロジェクト MMM のメンバー諸氏には、本研究に関する様々な議論をして頂きました。深い感謝の念を表します。

Bibliography

[Yan94] Tak W. Yan, Hector Garcia-Molina: SIFT - A Tool for Wide-Area Information Dissemination, 1994.
ftp://db.stanford.edu/pub/sift/sift.ps

[Ohk93] Masaaki Ohkubo, Naoki Kobayashi, and Toru Nakagawa: Design of an Information Skimming Space, *Proceedings of the ACM Multimedia 93*, pp. 365-371, Aug 1993.

[Bel92] Nicholas J. Belkin: Information Filtering and Information Retrieval: Two Sides of the Same coin?, *Communications of the ACM*, Vol. 35, No. 12, pp. 29-37, Dec 1992.

[Mor94] 森田 昌宏: 情報洪水の緩和のための情報フィルタリングの実現, *JAIN Symposium*, Jan 1994.
http://www.jaist.ac.jp/jaist/is/labs/sdhi-nlcab/papers/home-j.html

[Mor93] 森田 昌宏: 情報フィルタリング技術の現状と展望, 電子情報通信学会技術報告 (人工知能と知識処理), Vol. 93, No. 153, July 1993.
http://www.jaist.ac.jp/jaist/is/labs/sdhi-nlcab/papers/home-j.html

[Loe92] Shoshana Loeb: Architeching Personalized Delivery of Mutimedia Information, *Communications of the ACM*, Vol. 35, No. 12, pp. 39-47, Dec 1992.

[Mur95] 村上 列: WAIS による情報発信, *Software Design*, pp. 33-40, Feb 1995.

[Hat94] 服部 和子: 新聞紙面上の見出し地紋における視覚実験的検討, 「真ん中見てください」慶応義塾大学福田研究室, vol. 3, pp393-416, 1994.

[Par91] K. Parsaye, M. Chignell, S. Khoshafian, H. Wong, 近谷 英昭 訳: 知的データベース, オーム社, 1992.

[Fur86] Furnas, G. W.: Generalized fisheye views, *CHI'86 Proceedings*, Boston, MA, pp. 16-23

[Bac92] Paul. E. Baclace: Competitive Agents for Information Filtering, *Communications of the ACM*, Vol. 35, No. 12, pp. 50-60, Dec 1992.

[Del93] L. Delgrossi, C. Halstrick, D. Hehmann, R. G. Herrtwich, O. Krone, J. Sandvoss, C. Vogt: Media Scaling for Audiovisual Communication with the Heidelberg Transport System *Proceedings of the ACM Multimedia 93*, pp. 99-104, Aug 1993.