人　工　知　能　100－6
（1995．5．24）

スペクトル理論による情報フィルタ

下郡 信宏　　月本 洋
(株) 東芝 研究開発センター

新着の文書群から利用者にとって不要と思われる文書を自動的に捨ててしまう、情報フィルタについて述べる。様々な情報フィルタが存在するが、提示した文書に対して利用者からのフィードバックを受け利用者の好みを学習する情報フィルタを扱う。フィルタリングする文書をキーワードの集合で表現すると各キーワードが属性、キーワードが文書に出現「する／しない」が属性値、利用者に提示「する／しない」をクラスとみなす事が可能であり、単純なクラス判定問題となる。本稿では情報フィルタに適していると思われるスペクトル理論を適用した情報フィルタを試作し、実験を行った結果について述べる。

An Information Filter using Spectural Techniques

Nobuhiro Shimogori, Hiroshi Tsukimoto
Research & Development Center, Toshiba Corporation

This paper presents a system for automaticaly filtering out unneeded information, usually called an information filter. Among various information filters we deal with systems which learn on user's feedback. If we represent the information to be filtered with its keywords, the system can be seen as an ordinary classification system: with keyword as an attribute, the existence of the keyword within the information as the value, whether or not to show the information to the user as the class. We used a learning algorithm called Spectural Techniques for our system. Results of a filtering experiment is shown.

1 はじめに

計算機の普及により電子的に流通する情報が増え、またネットワークの発達により、誰もが情報の発信元となれる様になり情報氾濫がますます現実のものとなりつつある。例えば、ネットワークニュースなどでも利用者が得たい情報が膨大なゴミ情報の中に埋もれてしまい、タイムリーに必要な情報が得られないなどの問題が生じている。この様な状況において利用者の好みの情報を学習しながら、自動的に不要な情報を捨てる情報フィルタの必要性が増している。本研究では、判定すべき文書に含まれるキーワードの組み合わせを学習する事により利用者の興味の範囲を特定する情報フィルタについて述べる。

2 情報フィルタ

文書の内容を表現する何らかの特徴となる要素(例えばキーワード)を取り出し、それら特徴の組合せとして文書を近似する事が出来れば、情報フィルタは、単純なクラス判定問題と同じと考えられる。即ち文書をキーワードの集合とみなせば、情報フィルタが扱う全キーワードが属性の集合となり、キーワードがフィルタリング中の文書に含まれるか否かが属性値となる。更にその文書を利用者に提示する必要があるか否かがクラスとなる。例えば、キーワードが keyword0 ～ keyword9 の10個のであった場合、フィルタリング中の文書に keyword2,keyword4,keyword6 が含まれていると、(0,1,0,1,0,1,0,0,0) という問題事例が生成さる。「利用者に提示するべき」と予測した場合には情報を利用者に提示し、利用者からのフィードバックを待つ。利用者からのフィードバックとしては、単純に利用者が「良い／悪い」を入力する方式や、利用者が文書を表示していた時間などを用いる事が出来る。利用者からのフィードバックは問題事例のクラスとなりこれを教師信号として学習を行う。この様に考えると情報フィルタはクラス判定問題の単純なアプリケーションである。利用者の好みを学習する情報フィルタとしては、ニューラルネットを用いた情報フィルタ [1] や Latent Semantic Indexing(LSI) を用いた情報フィルタ [2] が存在する。

3 スペクトル理論とその適用

今回作成した情報フィルタでは、予測／学習に Linial らのアルゴリズム (スペクトル理論と呼ぶ)[3] を使用した。以下に簡単にスペクトル理論の予測／学習方法を示す。

学習方法

X: 事例ベクトルの集合
X_i:X の中の i 番目の事例ベクトル
事例ベクトルは、$X_i = (x_1, x_2, ...x_n)$ の様に記述し、j 番目の属性値が真の場合には $x_j = 1$ となり、偽の場合には $x_j = 0$ となる。
属性の組合せの全体を S で表し、例えば属性数が 3 の場合には、
$S = ((1),(2),(3),(1,2),(1,3),(2,3),(1,2,3))$
となる。S の要素は s で表す。
$\chi_s(X_i)$ は $\sum_{j \in S} x_j$ が偶数なら 1 を、奇数なら -1 をその値とする関数。
$f(X_i)$ は i 番目の事例に対するクラス。(-1,1) のいずれかの値を取る。

$$\alpha_s = \frac{\sum_{i=1}^{m} f(X_i)\chi_s(X_i)}{m}$$

上記の式により各 s に関して α を学習し、予測の際にもこの α_s を用いて予測を行う。

予測方法

問題事例 Y に対する予測値 $\overline{f(Y)}$ は上記の学習フェーズで学習した α を用いて、以下の式により求める。
$\overline{f(Y)} = Sign(\sum_{|s| \geq k} \alpha_s \chi_s(Y))$
ここで、$Sign(a)$ は $a > 0$ ならば 1 を、$a < 0$ ならば -1 となる様な関数である。$|s|$ とは、s に含まれる属性の数を表し次元と呼ぶ。最大の次元までを用いて学習／予測を行うと、非常に良い結果が得られるが、一定の次元で学習／予測を打ち切ってもある程度の精度が得られる事が分かっている。

適用

上記の予測／学習方法を情報フィルタに適用すると、以下の二つの問題が生ずる。

- 計算量
- 入力事例の偏り

これらは、情報フィルタで現れる事例の特性により起こる問題である。情報フィルタで扱う属性の数はキーワードの数と同じである為、数千から万に及んでしまう為に、計算量が膨大になってしまう。また、全キーワードの内、一つの文書の中に出現するものは、極一部に過ぎない。即ち、事例ベクトルを見た場合、ほとんどの属性値が0であり、1%にも満たない程度に1の属性値が存在する様な事例となる。更に、一般に情報フィルタは大多数のゴミ情報の中から、小数の有益な情報を拾い出すシステムである為に、ほとんどのクラスが-1であるという問題もある。この為に、Linialらのアルゴリズムをそのまま適用しても、全ての予測結果が-1になってしまい正しいフィルタリングが行えない。

システム

作成したシステムではほとんどの属性値が0である事に着目して高速化を行った。ほとんどの属性値が0であるという事は、ほとんどのsにおいて、$\chi_s(X) = 1$である事を意味する。従って、全ての属性値が0であった場合の予測値$f(0)$を予め計算しておき、ここから$\chi_s(X) = -1$となる様な、α_sに関してのみ修正する事により予測値を求める。今回の実験で使用したデータの場合、全属性数は約700であり、一つの文書に含まれるキーワードは平均3.5個である為、例えば2次でsの組み合わせを求めた場合、通常に求めると、${}_{700}C_2 = 244650$回の計算を行わなければならないが、$\chi_s(X) = -1$となるものは$3.5 * 696.5 = 2437.75$回であり、これらについてのみ計算すれば良い事になる。同様に学習フェーズでは、α_sは$\chi_s(X_i)$に依存するために、ほとんど場合に1であるならば、ほとんどのα_sに$f(X)$を加える事を意味する。従って$\sum f(X)$をα_sとは別に記憶しておき(これを、α_0と呼ぶ)、α_0との差分を個々に記憶する方法を用いれば、予測段階と同様の方法で計算量を減らす事が可能となる。更に予め全てのキーワードテーブルを用意するのは不経済な為、キーワードテーブルに登録されていないキーワードが出現する度に学習フェーズで登録する方法を取った。新規のキーワードは今までに学習に用いた事例の中に属性としては存在したが、属性値は全て偽であったと考える事が出来る、即ち$\alpha_{(n)}$を求める際の、$\chi_n(X)$は常に1であった為に、nを新しいキーワードの属性番号とすると、$\alpha_n = \alpha_0$である。同様に、$\alpha_{(i,..,j,n)}$は、新規のキーワードが登場するまでは、一度もキーワードnは出現していない為、その値は次元の一つ下の$\alpha(i,..,j)$と同じである。以上の様にすれば新規のキーワードが出現した時点での追加登録が可能である。情報フィルタが学習途上にある場合には、ほとんどのキーワードを知らない可能性があり、特に最初のフィルタリングセッションでは、何も学習していない為に予測は不可能である。そこで、予測に使用するキーワードの内、新規なキーワードの割合が一定以上を越えてしまった場合には、取り敢えず利用者に提示する方式を取っている。

次に、事例のクラスは文書を利用者に提示しない事を意味する-1が全体の約9割を占めている。従って学習が進むと共に、全ての問題に対して-1を予測する様になってしまう。この原因はクラスが-1の事例を多く学習すると、$f(0)$が負の大きな値を持ってしまい、α_sを加えても$Sign$の引数が負の値のままである為に起こる。これを解決する為に、$Sign$による閾値を設けずに、以下の式で求まる$\overline{f(Y)'}$を文書を利用者に提示するかどうかの確信度として利用し、フィルタリングを行った文書群を確信度の高い順に並べ換えるシステムとした。

$$\overline{f(Y)'} = \sum_{|s| \geq k} \alpha_s \chi_s(Y)$$

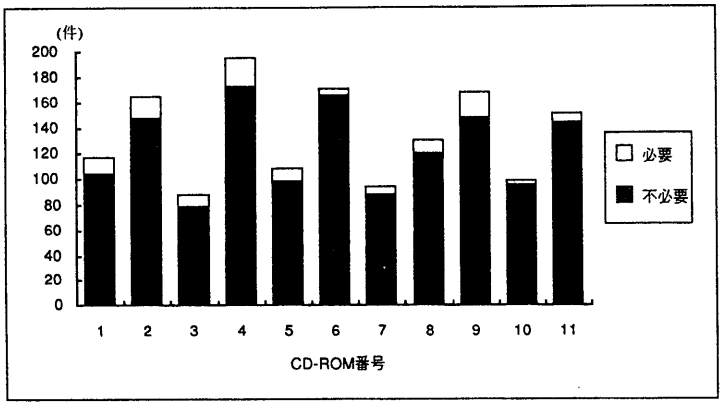


図 1: データの件数

これに伴い、同時にフィルタリングした文書の相対的な確信度にしか意味が無くなってしまい、全ての確信度に加えられる f(0) を求める必要も無くなった。

4 実験

特許公報をフィルタリングする実験を行った。特許公報を使用した理由は以下の通り。

- 1994 年より、特許公報は電子的な文字列として入手可能である
- 情報が氾濫している分野である
- 利用者に真剣に利用して貰える
- 文書の質が比較的揃っている

特に、特許文はネットワークニュースの記事などとは異なり、発明の内容を表すという同一の目的の元に書かれている為に、ネットワークのニュースの様に情報提供が目的で始まったスレッドが途中から喧嘩になってしまうなどという難しさが無く、フィルタリングの対象としては適している。また、特許の氾濫は実際に利用者が困っている分野であり、全てを読む事を考えるとシステムが学習するまで利用者が辛抱強く育てて貰えるメリットもある。

今回の実験では、1994 年の 1 月〜5 月に公開された特許公報から被験者に関連のある特許公報を選び出す実験を行った。使用した特許公報は CD-ROM で提供されており、特許庁から公布されている全分野の公開公報 CD-ROM ではなく、財団法人 日本特許情報機構が発行する分野別公開公報の物理分野を使用した。被験者は報告者一人。特許には分類 (IPC) が細かく付与されており、キーワードを自動的に抽出するとその能力により学習結果が左右されてしまう為、文書を表現するキーワードには IPC コードを使用した。各特許には平均 3.5 個の分類が付加されていた。実験の手順を以下に示す。

1. 特許分類を用いて予め大まかなフィルタリング対象を絞り込む
2. 絞り込んだ結果得られた全ての特許公報を被験者が読み、必要／不必要の判定を行ない評価値 (クラス) として付加する
3. CD-ROM 1 枚分の対象公報をフィルタリングを行う
4. 利用者に提示すると判定した公報の評価値を読み込み、学習を行う
5. ステップ 3 に戻る

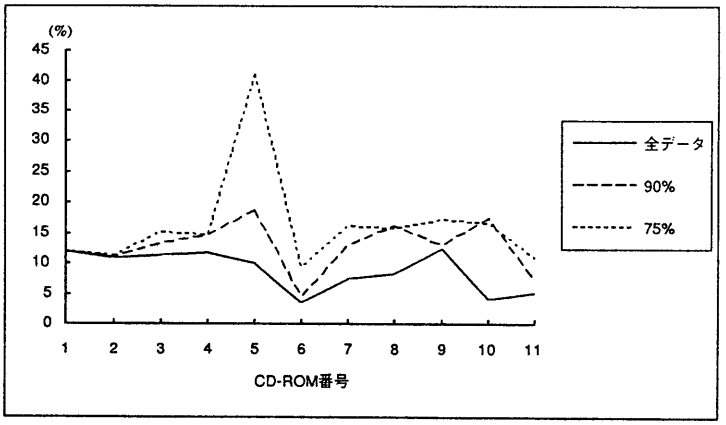
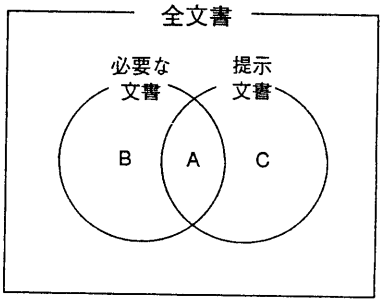


図 2: 精度

被験者が評価した文書は全部で1485件であり、そのうち被験者が必要としてる文書は133件であった。データの分布を図1に示す。CD-ROMにより、評価した文書の数も88～195件と大きく異なり、被験者が必要と判定した文書の数も4～23と大きく異なった。図1からも分かる通り、例えば10番目のCD-ROMでは、98件中4件しか必要な文書が含まれていないものもあり、ばらつきが大きい。情報フィルタの性能を類似検索に習って、再生率(recall rate)と精度(precision rate)を用いて示す。再生率とはフィルタリング対象の母集団に含まれていた必要な文書のうち、実際に利用者に提示されたものの割合を意味し、精度とは、利用者に提示された文書の中の必要な文書の割合を意味する。つまり以下の図では $\frac{A}{A+B}$ を再生率、$\frac{A}{A+C}$ を精度と呼ぶ。一般的に、再生率を上げると判定の難しい文書は提示してしまう傾向が強くなる為に精度は下り、逆に精度を上げると判定の難しい文書は提示しない方向に働く為に再生率は下がるというトレードオフの関係にある。今回作成したシステムは文書の提示／非提示を判定するのでは無く、重要と思われる順に文書を並べ換える為に、利用者が付加した評価値を用いて、再生率が特定の値になるまで順に提示し、提示した文書で精度を求めた。これにより、再生率を一定に保ちながら、精度の学習を見る事が可能となる。また、再生率 90％とは90％以上を指しており、必要な文書が9件しかない場合には、8件では90％に満たない為、9件全てを提示するまで提示対象を広げている。図2は全文書、再生率=90％および、再生率=75％における精度を表している。右に行くに従って、学習が進んだ状態となっている。値は高い程、無駄な情報を提示しなかった事を意味する。CD-ROM番号5だけは異常に精度が良いが、これは偶然である。それ以降の精度は15％前後で安定している。図3は全文書の内、再生率=90％、再生率=75％の場合の提示率を表している。提示率とは、フィルタリング前の文書から利用者に提示した文書の割合である。先程の図では、$\frac{A+C}{\text{全文書}}$ にあたる。この結果、学習が進んだ段階では、全体の約半分の情報を見れば必要としている情報の90％が得られる事が分かる。



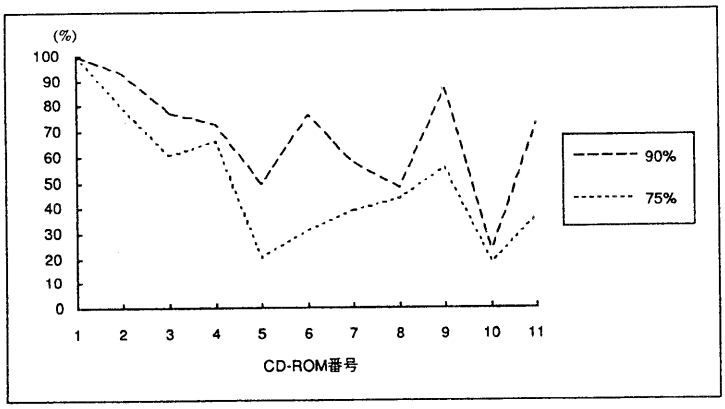


図 3: 提示率

5 今後の課題

今回使用した特許公報のデータをACT*[4]を用いた情報フィルタ[5]で同様にフィルタリングさせた。ACT*を用いた情報フィルタでは、予測方式が異なる為に、今回の様に再生率を固定する事が出来ない為に、単純に比較する事は出来ない。しかし、同程度の再生率で、80%程度の提示率であった事から、ACT*を用いた情報フィルタよりも確実にそのフィルタリング能力は向上している。また、提示率に波はあるものの学習と共に確実に下がっており、学習が進めば安定するものと思われる。計算に要する時間もCD-ROM一枚辺り約1分で終了しており、現実的な計算量で収まっている。今回のフィルタリング実験ではsを2次で学習／予測を行ったが、次数を上れば更に精度の良いフィルタリングが行えると期待出来る。しかし計算量が膨大になり、今回の実験では計算が終了しなかった。また、記憶している属性数の増加と共に、計算量及び学習結果の記憶量が増加してしまう為に、今後は不要な属性の削除方法や、計算量を更に削減する方法を検討しなければならない。

参考文献

[1] A. Jennings: A User Model Neural Network for Personal News Service,*User Modeling and User-Adapted Interaction* , Vol 3, No.1, pp.1-25,1993.

[2] Peter W. Foltz and Susan T. Dumais: Personalized Information Delivery: An Analysis of Information Filtering Methods,*CACM* , Vol 35, No.12, pp51-60, 1992.

[3] N. Linial, Y. Mansour, and N.Nisan: Constant Depth Circuits, Fourier Transform, and Learnability, *Journal of the Association for Computing Machinery* , Vol 40, No.3, pp.607-620, 1993.

[4] John R. Anderson: The Architecture of Cognition,*Harvard University Press Cambridge, Mass.*, 1983.

[5] 下郡 信宏、月本 洋: 利用者モデルの構想: 情報フィルタによる利用者情報の収集, 情報処理学会第 *49* 回全国大会, pp.5103-5104